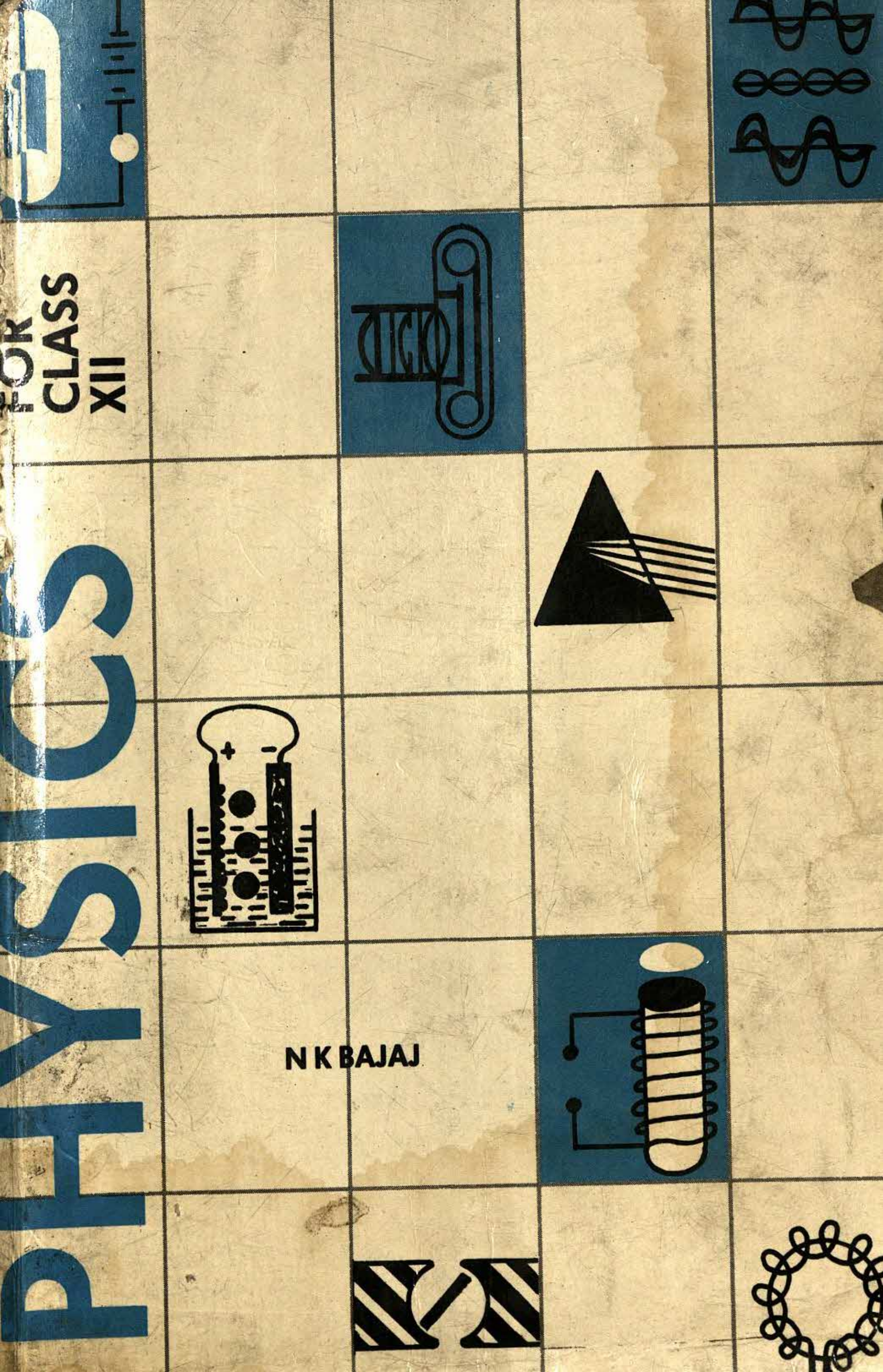
PHYSICS

FOR CLASS XII

N K BAJAJ



PHYSICS
FOR CLASS XII

PHYSICS
FOR CLASS XII

N K Bajaj
St Stephen's College
University of Delhi
Delhi

S. C. E. R. T. West B
Library
Calcutta

Tata McGraw-Hill Publishing Company Limited
NEW DELHI

McGraw-Hill Offices
New Delhi New York St Louis San Francisco Auckland Bogotá Guatemala
Hamburg Lisbon London Madrid Mexico Milan Montreal Panama
Paris San Juan São Paulo Singapore Sydney Tokyo Toronto

I.C.E.R.T. W.B. LIBRARY
Date 20.1.99
Accn. No. 9806

© 1990, TATA McGRAW-HILL PUBLISHING COMPANY LIMITED

No part of this publication can be reproduced in any form or by any means without the prior written permission of the publishers

This edition can be exported from India only by the publishers, Tata McGraw-Hill Publishing Company Limited

Assistant Sponsoring Editor: AMIT BHATTACHARJEE
Assistant Production Editor: KAUSHIK GHOSH
Production Officer: B L DOGRA

Published by Tata McGraw-Hill Publishing Company Limited,
4/12 Asaf Ali Road, New Delhi 110 002, and printed at
Rajkamal Electric Press, B 35/9 G T Karnal Road, Delhi 110 033

PREFACE

This textbook is written strictly in accordance with the latest physics syllabus for class XII prescribed by the Central Board of Secondary Education, New Delhi, under the 10 + 2 + 3 pattern of education.

The National Policy on Education, adopted in 1986, lays considerable emphasis on student-oriented rather than teacher-oriented books. According to this policy, books must be designed in such a way that a student is able to read and understand on his own without undue reliance on formal classroom instruction. The student must play an active role in the learning process and the teacher a passive one, serving only as a guide. Merely reading about the experiences and observations of others is not stimulating. If students are to understand and enjoy physics, they must have some of these experiences themselves. To encourage this kind of direct experiences with reality, I have recommended easily performed "home" experiments in the main body of the text. The student is advised to perform these experiments which require easily available material.

The objective of this book is to present physics in an interesting, understandable and enjoyable manner. An appreciation, rather than a working knowledge of physics is intended to be inculcated. Therefore, stress has been laid on the concepts of physics, rather than the mathematical aspects or techniques. Although calculus has been used, the mathematics involved is quite simple. The physical meaning of the mathematics used is clearly explained, thus enabling the student to appreciate the important part mathematics plays in the development of physics. Stress has been laid on practical applications throughout the book. The language is kept as simple as possible.

The coverage of topics is more or less traditional, but I have departed from the conventional approach in which general principles or laws are stated first and various results are deduced from them. To introduce a physical concept, I have begun with commonplace observations of our daily experience and cited concrete examples and then proceeded to suggest generalizations or laws that unify these observations. A study of the history of science reveals that this approach is more in keeping with how science is actually practised.

The introduction of each new concept is followed by a number of solved numerical examples. The book contains 200 solved examples.

It is my earnest hope that these examples will enable students to master a concept or a technique before they move on to the next. The physical concepts are amply illustrated by means of diagrams and graphs wherever possible. The book contains over 450 illustrations.

I am conscious of the fact that some students can learn much more rapidly and can assimilate more difficult concepts, as compared to the average students. These better and brighter students often do not find a subject challenging enough. To sustain and stimulate the interest of such students, some thought-provoking topics are discussed as supplements at the end of each chapter. The average student may omit these supplements at first reading; this will not affect his or her learning process.

The questions at the end of each chapter have been carefully designed. The numerical problems have been graded in order of the level of difficulty and the answers to all problems are given at the end of the book. The book contains over 300 numerical problems. For a quick recapitulation, a summary of the salient features is included at the end of each chapter.

No scientific theory is born in a vacuum. Each scientist builds on the work of his predecessors or contemporaries. Therefore, a brief historical development of ideas has been included in the body of the text wherever necessary.

I wish to express my deep sense of gratitude to my colleagues and students whose encouragement and cooperation have been a source of inspiration. It is a pleasure to thank Mr. Vineet Bajaj for going through the solutions of the numerical problems and making helpful suggestions and comments. Finally, I would like to thank my wife Mrs. Kamlesh Bajaj whose dedication and encouragement made this book possible.

Suggestions for the improvement of the book would be most welcome.

N K BAJAJ

A NOTE TO THE STUDENT

The main objective of this book is to present physics in an interesting, understandable and enjoyable manner so that you can understand the physical phenomena of the world around you. Each phenomenon is very carefully described by citing concrete examples from our everyday experience and a rational scientific explanation is provided for each phenomenon starting from fundamental principles. It is my earnest hope that a careful reading of the text in each chapter, and going through solved examples and working out the problems at the end of each chapter, will contribute immensely towards your understanding of the finer points of the subject of physics.

The somewhat bulky size of the book should not scare you. The book has assumed this size because of its following salient features:

1. The detailed steps of all mathematical derivations are given, starting from fundamental principles and bringing out clearly the mathematical logic and technique involved.
2. The introduction of each new concept is followed by a number of solved numerical examples. The book contains as many as 181 solved examples which will enable you to master a concept before you move on to the next. The examples have been carefully graded. We begin with simpler examples and gradually go on to more difficult ones involving the use of more than one concept or one formula.
3. The physical concepts are illustrated by means of a large number of diagrams and graphs. The book contains about 295 illustrations.
4. To sustain the interest of brighter students, supplements are appended at the end of a chapter which deal with more challenging and thought-provoking topics. The average student may omit these supplements at first reading.

The book contains about 400 questions and numerical problems. The numerical problems have been graded in order of the level of difficulty and the answers to all problems are given at the end of the book. You are strongly urged to solve as many problems as you can. Do not despair if you cannot solve all of them; some are quite difficult. If you are unable to solve a particular problem, read the relevant text more closely, especially the solved examples. You should develop the habit of working out at least 2 or 3 problems *regularly*. This is of the *utmost* importance. Only then will you be able to enjoy, appreciate and comprehend the subject of physics. Problem-solving not only sharpens your intellect and intuitions but can also be a source of immense fun and satisfaction. So keep solving problems!

N K BAJAJ

CONTENTS

Preface v

A Note to the Student vii

1 Electrostatic Force and Electrostatic Field **1**

1.1 Introduction: Frictional electricity *1*
1.2 Two kinds of electric charge *1*
1.3 Quantization of electric charge *3*
1.4 Conservation of electric charge *4*
1.5 Electric charge originates in the atom *5*
1.6 Coulomb's law of electric force *6*
1.7 Electrostatic induction *9*
1.8 Conductors and insulators *10*
1.9 Two methods of charging conductors *11*
1.10 Forces among many charges: the superposition principle *16*
1.11 The electric field: A useful concept *25*
1.12 The electric dipole *37*
1.13 Electric lines of force *41*
Supplement *45*

2 Electrostatic Potential and Capacitance **53**

2.1 Introduction *53*
2.2 An important property of electrostatic fields *53*
2.3 Electrostatic potential and potential difference *57*
2.4 Potential due to some charge distributions *61*
2.5 From electric potential to electric field *76*
2.6 Gauss's theorem for electrostatics *81*
2.7 Conductors and insulators *91*
2.8 Capacitors and capacitance *97*
2.9 Dielectrics and polarization of matter *113*
2.10 Atmospheric electricity *123*
2.11 The electrostatic generator *126*
Supplements *128*

3 Current Electricity **140**
3.1 Introduction *140*
3.2 Potential difference and current *140*
3.3 Mechanism of flow of current in a metallic conductor *143*
3.4 Ohm's law: resistance and resistivity *150*
3.5 Internal resistance *161*
3.6 Cells in series and in parallel *163*
3.7 Resistances in series and in parallel *164*
3.8 Kirchhoff's rules *170*
3.9 Measurement of current, potential difference and resistance *178*

4 Thermal and Chemical Effects of Currents **201**
4.1 Introduction *201*
4.2 Heating effect of current: Joule's law *201*
4.3 Chemical effect of current *211*
4.4 Electrochemical cells *222*
4.5 Thermoelectricity *231*
Supplement *244*

5 Magnetic Effects of Current **253**
5.1 Introduction *253*
5.2 Oersted's experiment *253*
5.3 Biot-Savart law and magnetic field due to a current *254*
5.4 Force on a moving charge in a magnetic field *265*
5.5 Lorentz force *272*
5.6 Ampere's circuital law *274*
5.7 Current loop and magnetic dipole *279*
5.8 Torque on a current loop and galvanometer *280*
5.9 Motion of a charged particle in a uniform magnetic field *286*
5.10 Motion of a charged particle in perpendicular magnetic and electric fields *296*
5.11 A solid in perpendicular electric and magnetic fields: The Hall effect *299*

6 Magnetism **308**
6.1 Introduction *308*
6.2 Magnetic dipole *309*

6.3 Magnetic field lines *312*
6.4 Potential energy of a magnetic dipole in an external magnetic field *314*
6.5 Magnetic analogue of Gauss's theorem in electrostatics: Gauss's theorem in magnetism *316*
6.6 The magnetic field of the earth *318*
Supplement *326*

7 Electromagnetic Induction **337**
7.1 Introduction: phenomenon of electromagnetic induction *337*
7.2 Magnetic flux *338*
7.3 Laws of electromagnetic induction *340*
7.4 Electromagnetic induction and Lorentz force: Motional EMF *344*
7.5 Methods of generating induced EMF *345*
7.6 Eddy currents *352*
7.7 Self and mutual inductance *355*
7.8 Electrical machines: AC generator and transformer *361*
Supplements *367*

8 Alternating Current Circuits **377**
8.1 Introduction *377*
8.2 Growth and decay of current in an inductor and a capacitor *378*
8.3 Some simple properties of complex number *385*
8.4 AC circuits with resistance, inductance and capacitance *387*
8.5 Energy and power associated with L, C and R *402*
8.6 Oscillations in an LC circuit *409*
8.7 Real resistors, capacitors and inductors *411*

9 Electromagnetic Waves **419**
9.1 Introduction *419*
9.2 Maxwell's displacement current *419*
9.3 Electromagnetic waves *428*
9.4 Experimental confirmation of electromagnetic waves *436*
9.5 Spectrum of electromagnetic radiation *438*

9.6 Electromagnetic radiation and earth's atmosphere *439*
Supplement *443*

10 Wave Optics **447**
10.1 Introduction *447*
10.2 Rays and wavefronts: Huygens' principle *447*
10.3 The laws of reflection and refraction *450*
10.4 Light—an electromagnetic wave *454*
10.5 Interference *457*
10.6 Colours of thin films *472*
10.7 Diffraction *476*
10.8 Polarization of light *487*
10.9 The doppler effect for light *494*
Supplements *497*

11 Ray Optics and Optical Instruments **504**
11.1 Introduction 504
11.2 Photometry *504*
11.3 Velocity of light *506*
11.4 Reflection of light from spherical mirrors *510*
11.5 Refraction of light *519*
11.6 Total internal reflection: critical angle *523*
11.7 Refraction in a prism *528*
11.8 Refraction at a spherical surface *532*
11.9 Refraction in a thin lens *535*
11.10 Combinations of thin lenses *541*
11.11 Dispersion of light *545*
11.12 Types of spectra *550*
11.13 Colour of the sky *553*
11.14 Defects in optical images *555*
11.15 Camera *560*
11.16 Microscope *561*
11.17 Telescope *564*
11.18 Resolving power *566*

12 Electrons and Photons **574**
12.1 Introduction *574*
12.2 Electrical discharge through a gas *574*
12.3 Cathode rays *576*
12.4 Charge to Mass Ratio of Electrons *578*

12.5 Millikan's Oil Drop Experiment for Measuring Electronic Charge *581*
12.6 Photoelectric Effect *585*
12.7 The Photo-Cell *591*
12.8 Compton scattering *593*
12.9 Wave nature of matter *594*

13 Atoms, Nuclei and Molecules **603**
13.1 Introduction *603*
13.2 Rutherford's model of atom *603*
13.3 Energy quantization: Bohr's model of hydrogen atom *606*
13.4 Atomic masses *617*
13.5 Discovery of the neutron *619*
13.6 Mass defect and binding energy *622*
13.7 Nuclear dimensions and nuclear density *626*
13.8 Nuclear forces *627*
13.9 Radioactivity *628*
13.10 Nuclear fission *639*
13.11 Nuclear fusion *643*
13.12 Molecules *645*

14 Solids and Semiconductor Devices **656**
14.1 Introduction *656*
14.2 Crystal structure *656*
14.3 Energy bands in solids *662*
14.4 Conductors, insulators and semiconductors *663*
14.5 Electrons and holes in semiconductors *665*
14.6 Intrinsic and extrinsic semiconductors *667*
14.7 Electrical resistivity of semiconductors *671*
14.8 PN junction *675*
14.9 PN junction diode *676*
14.10 The junction transistor *684*
14.11 Current gain in a transistor *688*
14.12 Current-voltage characteristics of a transistor *689*
14.13 Transistor as an amplifier *691*
14.14 Transistor as an oscillator *695*
14.15 Boolean operators *699*
14.16 Logic gates *703*
14.17 Combination of gates *706*
Supplement *709*

15 The Universe 717
15.1 Introduction *717*
15.2 Important constituents of the universe *719*
15.3 Astronomical instruments *719*
15.4 The Study of the Solar System *721*
15.5 The Sun *730*
15.6 The Stars *734*
15.7 The Milky Way *741*
15.8 Cosmological Theories of the Universe *743*

Appendix A 749

Appendix B 760

Answers 761

PHYSICS
FOR CLASS XII

1

ELECTROSTATIC FORCE AND ELECTROSTATIC FIELD

1.1 INTRODUCTION: FRICTIONAL ELECTRICITY

The word electricity is derived from the Greek word *elektron* which means *amber*. In the sixth century BC, a Greek philosopher Thales found that when a piece of amber (a kind of resin) is rubbed with fur, it acquires a remarkable property; it attracts small pieces of leaves, cork or dust. Today we know that many other materials, such as hard rubber, nylon, plastic, glass, sealing wax, etc. show the same effect. A substance which exhibits this effect is said to be *electrified* or *charged.*

You can see this kind of effect by doing a simple experiment. On a dry day, place a few tiny pieces of paper on a table. Take a plastic ruler or comb and rub it with dry hair or a piece of wool and bring it near the pieces of paper. You will observe that the pieces of paper are attracted towards the rubbed ruler or comb.

You may have experienced this attraction while combing your hair when dry or taking off a shirt made of synthetic fibre on a dry day. In each case the object is *charged* by the rubbing process (i.e. by friction) and it is said to possess *frictional* or *static electricity.*

No significant advance was made in the understanding of the science of electricity until about AD 1600 when William Gilbert (1540–1603), court physician to Queen Elizabeth I of England, undertook a detailed study of the kinds of material that would behave like amber. In his book *De Magnete* published in 1600, he gave the name *electrica* to substances like amber which when rubbed with suitable materials attracted tiny objects.

1.2 TWO KINDS OF ELECTRIC CHARGE

The next important step in the development of the science of electricity came about 100 years later when Du Fay showed that there are two kinds of charges. By rubbing various kinds of materials together, he was able to show that, under some conditions, they *repel* each other.

You too can show this effect. Rub a glass rod with silk and suspend it from a thread. Now bring another glass rod rubbed with silk near it as shown in Fig. 1.1. You will notice that the two glass rods *repel* each other. Similarly two ebonite rods rubbed with fur will also *repel* each other. But an ebonite rod rubbed with fur and a glass rod rubbed with silk will attract each other.

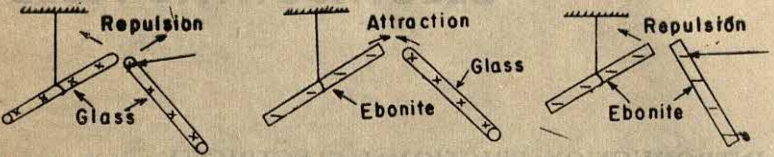


Fig. 1.1 *Like charges repel, unlike charges attract*

The above observations clearly suggest that the kind of charge developed on the glass rod when rubbed with silk must be opposite to the kind of charge developed on the ebonite rod when rubbed with fur. Benjamin Franklin (1706–1790), an American Physicist, suggested the convention that the charge on the glass rod when rubbed with silk be called *positive* and that on the ebonite rod when rubbed with fur be called *negative*. This convention is followed to this day. This choice was a historical accident and an unfortunate one since the electron (so fundamental and important to all the chemical and physical properties of matter in bulk) turns out to have a negative charge. We could as well have adopted the convention that the charge of an electron be taken to be positive. There is nothing intrinsically negative about the charge of an electron.

Glass and ebonite are not the only materials that can be charged by rubbing. In fact, any substance rubbed with any other substance, under suitable conditions, will become charged to some extent. The kind of charge developed on a substance can be found by testing it for repulsion when brought near a charged glass rod or a charged ebonite rod. When different materials are tested, it is found that they fall into the following two categories:

1. Some developed the same kind of charge as that developed on a glass rod when rubbed with silk. Conventionally, these materials are said to have acquired a *positive* charge, i.e. they are *positively charged.*
2. Others developed the same kind of charge as that developed on an ebonite rod when rubbed with fur. Conventionally, these materials are said to have acquired a *negative* charge, i.e. they are *negatively charged.*

These experiments verify a fundamental law of static electricity, namely, *like charges repel, unlike charges attract.*

One fundamental property of electric charge is that it exists only in two (not three or more) varieties, positive or negative. The experimental fact is that all charged bodies can be divided into just *two* classes such that all members of one class repel each other and attract members of the other class (see Table 1.1 below)

Table 1.1 Frictional electricity leading to two kinds of charge

Class I (Positive charge)	Class II (Negative charge)
Glass rod	Silk cloth
Fur cap	Ebonite rod
Woollen cloth	Plastic pen
Woollen carpet	Rubber shoes

Note: If we have three charged bodies A, B and C, then experiment shows that if A and B repel each other and if A attracts C then we always find that B attracts C. Thus like charges always repel and unlike charges always attract and not the other way round. Physicists do not have a definite answer to why this universal law prevails.

1.2 QUANTIZATION OF ELECTRIC CHARGE

In Franklin's time electric charge was thought of as a continuous fluid. The atomic theory of matter, however, has shown that fluids themselves, such as water and air, are not continuous but are made up of atoms. Experiments have shown that the electric charge is not continuous. Any charge, small or big, comes only in discrete 'lumps' or 'packets' of a certain minimum charge. We use the symbol e for this fundamental unit of charge which is the charge on a proton. The charge on an electron is $-e$. We shall state its numerical value later. The charge on a neutron is zero.

Any charged body, large or small, has a total charge which is some *integral* multiple of e: $\pm e$, $\pm 2e$, $\pm 3e$, . . . , $\pm ne$ where n is an integer. This experimental fact is called the *quantization of electric charge*. Thus the carbon nucleus, for example, always has a charge of $+6e$, never $5.972e$ or $6.013e$.*

---

*In recent years physicists have postulated the existence of particles of charge $1/3e$ and $2/3e$. These particles are named *quarks*. Till today such particles have not been detected in any experiment. However, if some day they are detected, that would not alter the fact that charge is quantized—it would merely reduce the size of the basic unit of charge.

When a physical property exists in discrete 'packets' rather than in continuous amounts, the property is said to be quantized. Electric charge is quantized. The existence of atoms and particles such as electrons and protons indicates that mass is also quantized. Many other physical quantities (such as energy and angular momentum) prove to be quantized when suitably examined on the atomic scale.

The fundamental quantum (or unit) of charge e is so small that the 'graininess' of charge does not show up in large-scale experiments, just as we do not realize that the air we breathe or the water we drink are made up of atoms.

Note: The fact that the charge on a proton is numerically equal (but opposite in sign) to that on an electron has been checked experimentally to an astonishingly high degree of accuracy of one part in 10^{20}. We do not yet understand why their charges should be equal to such a fantastically precise degree. It appears that the quantization of charge is a deep and universal law of nature.

The quantization of charge cannot be explained on the basis of classical electromagnetism. But it is a fact of nature and must always be borne in mind while dealing with electromagnetic phenomena at a microscopic level. Our study of electricity and magnetism will be restricted to large-scale (macroscopic) phenomena for which the quantization of charge can be ignored. We shall, therefore, treat the charged bodies simply as carriers of charge which can change continuously. If the amount of charge on a body is a *huge* integral multiple of the basic unit e, we can ignore the structure (or graininess) of the charge and imagine it to be continuous.

Electric Charge is Additive

Just as the total mass of an extended body is the sum of the masses of the particles of which the body is made, *the total electric charge of an extended body is the algebraic sum* (i.e. the sum taking into account the sign of the charges) *of charges located at different points of it.* Thus two equal positive charges $+q$ and $+q$ add up to $+2q$, two equal negative charges $-q$ and $-q$ add up to $-2q$, but two equal and opposite charges $+q$ and $-q$ add up to zero. In other words, *the electric charge is additive.*

1.4 CONSERVATION OF ELECTRIC CHARGE

When a glass rod is rubbed with silk, a positive charge appears on the rod. Measurements show that a negative charge of an equal magnitude appears on silk. This suggests that rubbing (or friction) does not create charge but merely transfers it from one object to the other. Thus charge can neither be created nor destroyed. The total electric charge of an isolated system cannot change; it is fixed for all time. This is the principle of conservation of charge. This principle has been subjected to close experimental scrutiny both for large-scale phenomena and for those at the atomic or even nuclear level; no exceptions have been found to date. Charge conservation is a *global* law; *the total charge of the universe is fixed for all time.*

We will come across examples of conservation of electric charge when we study electrolysis, chemical effects of electric current, radioactive decay and nuclear reactions in subsequent chapters of this book.

Note: We have stated two properties of electric charge, namely, the quantization and conservation of electric charge. There is a third interesting property of electric charge which is worth mentioning. The electric charge on a body does not depend on its speed. Einstein's special theory of relativity tells us that the mass of a body depends on its speed: in fact the mass of a body increases as its speed increases. But the charge of a body moving anywhere in the universe with any speed will remain constant.

1.5 ELECTRIC CHARGE ORIGINATES IN THE ATOM

According to the atomic theory, matter is made up of atoms. An atom is made up of a dense positively charged nucleus, surrounded by a cloud of electrons. The nucleus varies in radius from about 1×10^{-15} m for hydrogen to about 7×10^{-15} m for the heaviest atoms. The outer diameter of the electron cloud (i.e. the diameter of the atom) varies from about 1×10^{-10} m to about 3×10^{-10} m, about 10^5 times larger than the nuclear diameter.

The nucleus of an atom consists of positively charged protons and uncharged neutrons. As stated earlier, the amount of the positive charge on a proton is equal to the amount of the negative charge on an electron. In a normal atom, the number of protons is equal to the number of electrons, so that the atom as a whole is electrically neutral. This is an astonishing fact of nature. It is indeed remarkable that, in bulk matter, the positive and negative changes occur in *exactly* equal amounts to fantastic precision, so that their effects completely cancel. If this were not so, matter would be subjected to enormous forces. A potato, for example, would explode violently if the cancellation were incomplete by as little as one part in 10^{10}.

The electrons can be transferred from one body to another. If a body loses one or more electrons, the body as a whole is left with a positive charge. If a body gains electrons, it will then have an excess of electrons and hence, will have a negative charge. It must be clearly understood that a body cannot, by ordinary methods, lose or gain extra protons which remain inside the nuclei of the atoms and cannot move from place to place.

Explanation of Electrification by Rubbing (or Friction)

All the phenomena of static electricity can be explained in terms of a transfer of electrons. The electrons in some materials, such as glass, are held less tightly than those in others, such as silk. Thus, when a glass rod is rubbed with silk, some electrons are transferred from glass to silk. Consequently, the glass rod develops a positive charge (due to deficiency of electrons) and silk develops an equal negative charge (due to excess of electrons). Similarly, the electrons in ebonite are more tightly bound than those in fur. Thus, when an ebonite rod is rubbed with fur, electrons are transferred from fur to ebonite. Hence, the ebonite rod acquires a negative charge and fur, an equal and opposite positive charge. Thus while the electric charge is indeed conserved, it can certainly be *transferred* from one body to another, basically by transferring electrons.

Note: It may be mentioned that the frictional forces are electrical in origin. The only way to pull an electron from an atom is to make a sufficiently strong electric force act on it. Friction provides this force.

1.6 COULOMB'S LAW OF ELECTRIC FORCE

The first quantitative measurements of the magnitude of electric forces between charged bodies were made by a French scientist, Charles Augustin de Coulomb in the year 1785. His apparatus, known as Coulomb's torsion balance, shown in Fig. 1.2, resembles the hanging rod of Fig. 1.1, except that the charges in Fig. 1.2 are confined to small spheres A and B. Sphere B is fixed in position and Sphere A is movable as shown.

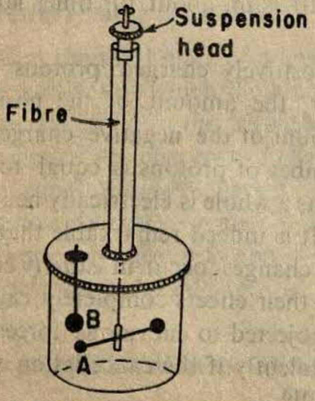


Fig. 1.2 *Coulomb's torsion balance*

When A and B are charged, the electric force on A will make it move thus twisting the suspension fibre. Coulomb cancelled out the twist by turning the suspension head through an angle θ required to keep the sphere at a particular known distance apart. The value of θ then gives a measure of the electric force acting on A. He repeated the experiment with different charges on A and B and for different known distances between them.

On the basis of his measurements, Coulomb arrived at a law, known after his name as Coulomb's law, which states that *the magnitude of the electric force between two charges is directly proportional to the product of the magnitudes of the charges and inversely proportional to the square of the distance between them*, i.e.

$$F \propto \frac{q_1 q_2}{r^2}$$

or

$$F = k\frac{q_1 q_2}{r^2} \tag{1.1}$$

where F is the magnitude of the electric force, which each charge exerts on the other q_1 and q_2 are the magnitudes of the charges, r is the distance between them and k is a constant of proportionality. Coulomb's law holds for charged bodies whose sizes are much smaller than the distance between them. Strictly the law holds only for point charges.

Coulomb's law (1.1) resembles the law of gravitation

$$F = G\frac{m_1 m_2}{r^2}$$

which was discovered by Newton about 100 years earlier. The charge plays the same role for electric force as mass does for gravitational force. These two forces are not related but are entirely independent phenomena. The gravitational force is always attractive but the electric force is attractive if $q_1 q_2 < 0$ (i.e. unlike charges) and repulsive if $q_1 q_2 > 0$ (i.e. like charges). This corresponds to the fact that there are two kinds of charges but apparently only one kind of mass.

Coulomb's Law in Vector Form

Suppose two bodies labelled 1 and 2 having charges q_1 and q_2 respectively are separated by a distance r. We assume that the bodies are much smaller in linear dimension than the distance r and are situated in a vacuum. We have seen that if $q_1 q_2 > 0$, they repel each other and if $q_1 q_2 < 0$, they attract each other (Fig. 1.3). In vector form, Coulomb's law is written as follows:

$$\mathbf{F}_{21} = \text{force on 2 due to 1} = k\,\frac{q_1 q_2}{r^2}\cdot\hat{\mathbf{r}}_{12} \qquad (1.2a)$$

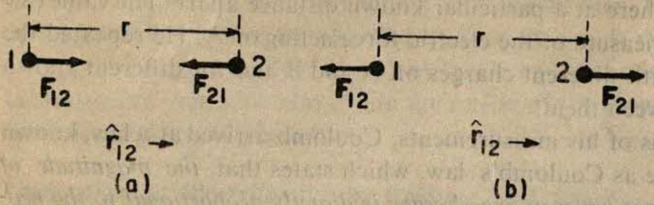


Fig. 1.3 *Coulomb's law: (a) attraction, (b) repulsion*

where $\hat{\mathbf{r}}_{12} = \mathbf{r}_{12}/r$, is the unit vector pointing from 1 to 2. Also

$$\mathbf{F}_{12} = \text{force on 1 due to 2} = k\,\frac{q_1 q_2}{r^2}\cdot\hat{\mathbf{r}}_{21} \qquad (1.2b)$$

where $\hat{\mathbf{r}}_{21}$ unit vector pointing from 2 to 1. It is obvious that

$$\hat{\mathbf{r}}_{21} = -\hat{\mathbf{r}}_{12}$$

Thus $$\mathbf{F}_{21} = -\mathbf{F}_{12}$$

Notice that the forces exerted by the two charges on each other are equal and opposite, i.e. they obey Newton's third law of motion. The forces are parallel to the line joining the two charges.

The value of the undetermined constant k in Eq. (1.1) depends upon the system of units we decide to choose for F, r and q and is to be determined experimentally. You are familiar with the SI units of F and r, they are newton and metre respectively. It now remains to determine the units of k and q. There are two alternatives: (i) k is arbitrarily given some fixed numerical value and Eq. (1.1) is used to determine the unit of charge, or (ii) the unit of charge may be taken as some arbitrary value and then k determined by experiment. The *Système International* (SI), which will be used throughout this book, uses the second alternative. An arbitrary unit of charge called the *coulomb* (symbol C) is chosen for charge and the force F(in newtons or N) between two charges kept at a known distance r(in metres or m) is experimentally measured; then Eq. (1.1) is used to determine k. The experimentally measured value is

$$k = 8.9875\times10^9\ \text{Nm}^2\ \text{C}^{-2}$$

For practical calculations we can take $k = 9\times10^9\ \text{Nm}^2\ \text{C}^{-2}$. Equation (1.1) gives us a definition of a coulomb of charge. If $q_1 = q_2 = 1$ coulomb, $r = 1$ metre, then $F = k$ newtons $= 9\times10^9$ N. Thus, *one coulomb is that quantity of charge that exerts a force of* 9×10^9 *newtons* *on an equal charge placed in vacuum at a distance of 1 metre from it.*

In the SI system k is written as $1/4\pi\epsilon_0$, where ϵ_0 is called the 'permittivity of vacuum' and its value is

$$\epsilon_0 = \frac{1}{4\pi k} = 8.854\times 10^{-12}\ \mathrm{C^2\ N^{-1}\ m^{-2}}$$

A more convenient name for the unit $\mathrm{C^2\ N^{-1}\ m^{-2}}$ is farad per metre ($\mathrm{F\ m^{-1}}$). Then Coulomb's law appears as

$$\mathbf{F}_{21} = \frac{1}{4\pi\epsilon_0}\cdot\frac{q_1q_2}{r^2}\hat{\mathbf{r}}_{12} \tag{1.3a}$$

Magnitude of force is

$$\mathbf{F} = \frac{1}{4\pi\epsilon_0}\cdot\frac{q_1q_2}{r^2} \tag{1.3b}$$

Although this complicates the expression for Coulomb's force, the factor 4π is deliberately introduced to simplify other equations that are used more extensively than Coulomb's law, as you will see later.

If the charges are situated in a medium other than vacuum, the quantity ϵ_0 in Eqs (1.3) is replaced by ϵ, the permittivity of the medium. The ratio $\epsilon_r = \epsilon/\epsilon_0$ is the *relative permittivity* which is also called the *dielectric constant* of the medium, the meaning of which will be explained in the next chapter. The value of ϵ_r for vacuum is unity by definition, for air $\epsilon_r \simeq 1$ and for water $\epsilon_r \simeq 81$.

The coulomb defined above is a very big unit of charge. Two charges of 1 C kept 1m apart in air will repel each other with a force of 9×10^9N which is an enormous force. The charge of an electron can now be expressed in terms of the coulomb. In the year 1917 R.A. Millikan (1868-1953) performed the famous oil-drop experiment for the determination of the charge on an electron. The accepted value is

$$\text{Charge on an electron} = -e = -1.6\times 10^{-19}\mathrm{C}.$$

which means that 6.2×10^{18} electrons together will have a negative charge of 1 C.

Note: Coulomb's law has been tested directly or indirectly over an enormous range of separations (r), from nuclear dimensions $r \simeq 10^{-15}$ m to several kilometres. It is found that the inverse square dependence on separation holds for more than 18 orders of magnitude in distance.

1.7 ELECTROSTATIC INDUCTION

At the beginning of this chapter we have stated that a plastic ruler when rubbed with hair or fur attracts pieces of paper. You may now wonder why a charged plastic ruler attracts a piece of paper that has no charge on it, because according to Coulomb's law, if one of the bodies has zero charge, the force on it must be zero. The atomic theory tells us how an attraction can exist in this case. When a negatively charged plastic ruler is brought near a neutral object, such as a piece of paper, the electrons in the object will feel a repulsive force. The electrons, will, therefore, move away from the end of the object nearer the plastic ruler to the opposite end which is farther from the ruler. Hence the nearer end of the object will have a positive charge (due to a deficit of electrons) and the farther end will have a negative charge (due to an excess of electrons) [Fig. 1.4(b)].

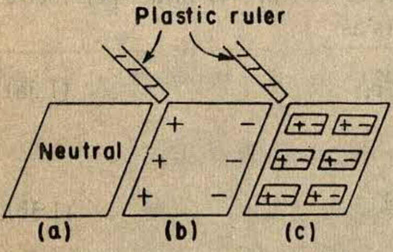


Fig. 1.4 *Electrostatic induction*

The object, as a whole, still has a net charge of zero, but there is a separation of charge. This charge separation is said to be induced by the charged ruler. Referring again to Fig. 1.4(b), we see that the negatively charged ruler must exert an attractive force on the positive end of the object and a repulsive force on the negative end. Because the positive end is closer to the ruler, Coulomb's law tells us that the attractive force must be greater than the repulsive force. Thus the net force on the object is attractive.

Some materials have practically no free electrons. Paper is an example of such material. How then is a piece of paper attracted by a charged ruler? The answer is that, although the electrons of paper atoms are not free to move in the body of the paper, they are still free to orient themselves within each atom. When a negatively charged ruler is brought near a piece of paper, charge separation occurs within each atom as shown in Fig. 1.4(c). The positive part of each atom will be closer to the ruler and again the net result is an attraction.

1.8 CONDUCTORS AND INSULATORS

We have seen above that electrons can migrate or travel within the body of a substance. The number of such electrons which are free to move is different for different substances. In nature, nearly all substances can be classified into two categories, namely, conductors and insulators.

Substances which have a large number of free electrons are called conductors of eletricity. They are called conductors because they can be used to carry or conduct electric charge from one place to another. Metals are good conductors of electricity. Silver is one of the best conductors. Examples of good conductors are silver, copper, aluminium, iron, mercury, coal, acids, alkalis, salts, earth, human body, etc.

On the other hand, *substances which have a negligibly small number of free electrons are called non-conductors or insulators*. They are very poor conductors of electricity. They are called insulators because they can be used to prevent charge from going where it is not wanted. Some common insulators are glass, wood, rubber, plastic, ebonite, mica, wax, paper, etc.

It is interesting that nearly all substances fall into these two categories, with very few exceptions which are neither good conductors nor insulators. These intermediate substances, such as germanium, carbon, silicon, etc. are called *semiconductors*. They play a very important role in solid state electronic devices.

1.9 TWO METHODS OF CHARGING CONDUCTORS

There are two methods by which an uncharged conductor can be charged, (a) by conduction and (b) by induction.

1. Charging by Conduction (or Contact)

A conductor can be charged by bringing it in contact with another charged conductor as shown in Fig. 1.5. The conductor B to be charged (say positively) is placed on an insulating stand. A positively charged conductor A (provided with an insulated handle) is made to touch the uncharged conductor B. Some positive charge of A will be transferred to B. Conductor A is now removed. Conductor B will be left with a positive charge.

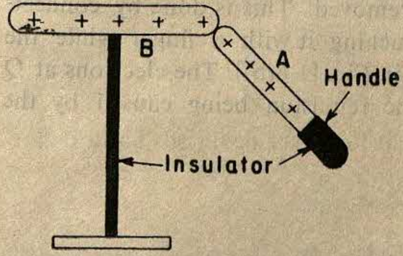


Fig. 1.5 *Charging by conduction (contact)*

This can be explained as follows. Since conductor A is positively charged, it has a deficiency of electrons. When it is brought in contact with a neutral conductor B, the free electrons of B will flow to A to make up for its deficiency. Due to a loss of electrons in B, it becomes positively charged.

On the other hand, if conductor B is to be negatively charged, a negatively charged conductor A is made to touch it. In this case the transfer of electrons takes place from A to B. Thus, in conduction, the charges are shared and the kind of charge developed is the same as that on the charging body.

It may be mentioned that insulators cannot be charged by conduction because they do not contain any free electrons. Hence no transfer of charge can take place in insulators.

2. Charging by Induction

Electrostatic induction may be used to charge an uncharged conductor. Figure 1.6 shows how a conductor B can be positively charged by induction. The conductor is mounted on an insulating stand and a negatively charged rod A is brought close to it (without touching it) [Fig. 1.6(a)]. The nearer end P of conductor B acquires a positive charge and the farther end Q a negative charge. This happens because the negatively charged rod A will repel some free electrons of B so that they are pushed from the nearer end P to the farther end Q. The nearer end, having lost some electrons, becomes positively charged while the farther end, the one that has gained those electrons, becomes negatively charged. In this process, no charge is transferred from the inducing body A to the induced body B, only a separation of charge takes place in B because of the presence of A. Hence if the rod A is removed, the negative charge at Q will be attracted by the positive charge at P. Therefore, the electrons will travel from Q to P, thus neutralising the charge on B which becomes uncharged. To prevent this, the electrons at Q must somehow be removed. This is done by connecting the end Q to the earth or by touching it with a finger while the rod A is held in place, as shown in Fig. [1.6(b)]. The electrons at Q are then repelled to the earth, the repulsion being caused by the

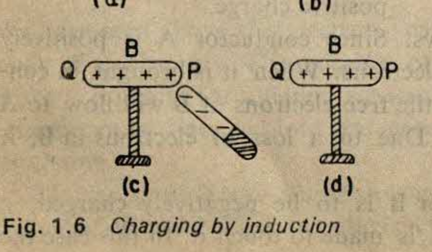


Fig. 1.6 *Charging by induction*

negative charge on A. The earth connection is now broken [Fig. 1.6 (c)] and the inducing rod A is then removed [Fig. 1.6(d)]. The positive charge at P now spreads in the conductor B which is, therefore, charged positively.

Thus to charge a conductor positively the inducing charge must be negative and vice versa. It may be mentioned that electrostatic induction may be used to obtain an almost unlimited charge from a single inducing charge. This is possible because, unlike charging by conduction, the charging by induction does not involve any transfer of charge.

EXAMPLE 1.1 An electron has a negative charge of 1.6×10^{-19} C and a mass of 9.1×10^{-31} kg. A proton has a positive charge of 1.6×10^{-19}C and a mass of 1.7×10^{-27} kg. Together they make up the hydrogen atom. Compare the electric force and the gravitational force between them. Given $\epsilon_0 = 8.85 \times 10^{-12} \mathrm{C^2N^{-1}m^{-2}}$ and $G = 6.67 \times 10^{-11}$ $\mathrm{Nm^2\,kg^{-2}}$.

Solution: The electric force of attraction between the electron and the proton is

$$F_e = \frac{1}{4\pi\epsilon_0} \cdot \frac{q_1 q_2}{r^2} = \frac{1}{4\pi\epsilon_0} \cdot \frac{e^2}{r^2}$$

The gravitational force between them is

$$F_g = G \cdot \frac{m_1 m_2}{r^2} = G \cdot \frac{m_e m_p}{r^2}$$

$$\therefore \quad \frac{F_e}{F_g} = \frac{e^2}{4\pi\epsilon_0} \cdot \frac{1}{G m_e m_p}$$

Substituting the given values of the various quantities we get

$$\frac{F_e}{F_g} \simeq 2.4 \times 10^{39}$$

Thus the electric force is 39 orders of magnitude stronger than the gravitational force. Hence, while considering electric force between bodies of ordinary masses, the gravitational force can always be ignored.

EXAMPLE 1.2 A polythene piece rubbed with wool is found to have a negative charge of 8×10^{-7}C. (a) Does wool acquire any charge in this process? If so, how much? (b) From which substance to which other are the electrons transferred? Estimate the number of electrons transferred. (c) Is there any transfer of mass in this process?

Solution: (a) Since the total charge is conserved, the wool will acquire an equal and opposite charge. Thus wool acquires a positive charge of 8×10^{-7}C.

(b) Since polythene acquires a negative charge due to rubbing with wool, it is obvious that electrons are transferred from wool to polythene. Since the magnitude of the charge on an electron is 1.6×10^{-19}C, the charge of 8×10^{-7}C corresponds to

$$\frac{8 \times 10^{-7}}{1.6 \times 10^{-19}} = 5 \times 10^{12} \text{ electrons}$$

Hence 5×10^{12} electrons are transferred from wool to polythene.

(c) Since the mass of an electron is finite but very small ($\sim 10^{-30}$ kg), there is a transfer of a finite (though negligibly small) mass from wool to polythene.

EXAMPLE 1.3 (a) Two insulated charged metallic spheres A and B have their centres separated by a distance of 50 cm in air. Calculate the force of electrostatic repulsion between them if the charge on each is $+5.0 \times 10^{-8}$ C. The radii of A and B are negligible compared to the distance of separation. (b) What is the force of repulsion if (i) the charge on each sphere is doubled and the distance between them is halved; and (ii) the two spheres are placed in water? Dielectric constant of water = 81.

Solution: (a)

$$q_1 = q_2 = +5.0 \times 10^{-8}\text{C} \qquad r = 50 \text{ cm} = 0.50 \text{ m}$$

The force of repulsion is given by

$$F = \frac{1}{4\pi\epsilon_0} \cdot \frac{q_1 q_2}{r^2} \tag{i}$$

where $\epsilon_0 = 8.85 \times 10^{-12}\text{C}^2\text{N}^{-1}\text{m}^{-2}$

Substituting the values of the various quantities we get

$$F = 9.0 \times 10^{-5} \text{ N}$$

(b) (i) If the charge on each sphere is doubled and the distance between them is halved, then Coulomb's law tells us that the force F would become 16 times, i.e.

$$F \simeq 1.4 \times 10^{-3} \text{ N}$$

(ii) If the spheres are immersed in water, the force between them is

$$F = \frac{1}{4\pi\epsilon} \cdot \frac{q_1 q_2}{r^2}$$

where ϵ is the permittivity of water. The dielectric constant (or relative permittivity) is

$$\epsilon_r = \frac{\epsilon}{\epsilon_0} \quad \text{or} \quad \epsilon = \epsilon_0\epsilon_r$$

$$\therefore \qquad F = \frac{1}{4\pi\epsilon_0\epsilon_r}\cdot\frac{q_1q_2}{r^2}.$$

It is given that $\epsilon_r = 81$. Hence if the two charged spheres are immersed or water (whose $\epsilon_r = 81$), the force of repulsion would decrease by a factor of 81. Hence

$$F_{\text{water}} = \frac{9\times10^{-5}}{81} = 1.1\times10^{-6}\text{ N}$$

EXAMPLE 1.4 Suppose the charged metallic spheres A and B in Example 1.3 above have identical sizes. A third metallic sphere of the same size, but uncharged, is brought in contact with sphere A and removed. It is then brought in contact with sphere B and removed. What is the new force of repulsion between A and B?

Solution: Let the charge of each sphere A and B be $+q$ which is given to be $+5.0\times10^{-8}$C. We have seen that the force of repulsion between them is $F = 9.0\times10^{-5}$ N.

When two identical metallic spheres are brought in contact, the charges on them are equalized due to the flow of free electrons. Thus when an uncharged sphere C is brought in contact with sphere A having a charge $+q$, and then removed, the total charge q is equally shared between the two so that the charge left on A is $+q/2$ and that developed on C is $+q/2$.

The sphere C carrying a charge $+q/2$ is now brought in contact with sphere B which is already carrying a charge $+q$. The total charge is $q/2 + q = +3q/2$ which must distribute equally on B and C. Thus when C is removed, B will have a charge of $+3q/4$ and C also has a charge of $+3q/4$. Hence when C is removed from both A and B,

$$\text{New charge on A} = +\frac{q}{2}$$

$$\text{New charge on B} = +\frac{3q}{4}$$

Since force is proportional to the product of the charges, it follows that the new force of repulsion between A and B is 3/8 of the earlier force (F). Hence, the new force of repulsion between A and B is

$$\frac{3}{8}\times9.0\times10^{-5} \simeq 3.4\times10^{-5}\text{ N}.$$

EXAMPLE 1.5 Two uncharged metal spheres A and B are supported on insulating stands and placed in contact as shown in Fig. 1.7. A negatively charged ebonite rod is placed near A but not in contact with it. Describe what happens when (a) the sphere B is momentarily touched with the finger (or earthed) and then the ebonite rod is removed, and (b) the ebonite rod is made to touch the sphere A.

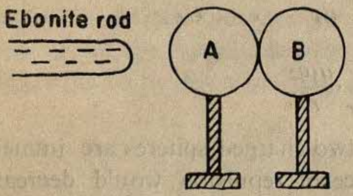


Fig. 1.7

Solution: (a) When a negatively charged ebonite rod is brought near A, it induces a positive charge on the side of A near the rod and a negative charge on the side of B remote from the rod. When the sphere B is earthed or touched with the finger, the negative charge on B flows to the earth due to the repulsion of the negative charge of the rod. If now the earth connection is broken and the inducing rod removed, the positive charge on A will spread to B so that both A and B have a positive charge.

(b) The charges on A and B remain unchanged if the ebonite rod is made to touch A since ebonite is an insulator and hence no charge can flow from it to the sphere A.

1.10 FORCES AMONG MANY CHARGES: THE SUPERPOSITION PRINCIPLE

Coulomb's law (1.3a) gives us the forces which two tiny (point) charges exert on each other when they are in vacuum and there is no other charge around them. How can we compute the force experienced by a charge if there are more than two charges? Suppose there are many point charges q_1, q_2, q_3, q_4, . . . situated at points P_1, P_2, P_3, P_4, . . . in space as shown in Fig. 1.8. We wish to find out the total force that acts on one of these charges, say, q_1 at P_1 due to all the other charges.

The answer to the problem is obtained by using the so-called *principle of superposition* which states that *the total force on a given charge is the vector sum of the individual forces exerted on it by all other charges, each individual force calculated from Coulomb's law*. The essence of the principle is that *the force between any two charges is completely unaffected by the presence of other charges*. This principle has been verified experimentally. Thus to find the *total* force on charge

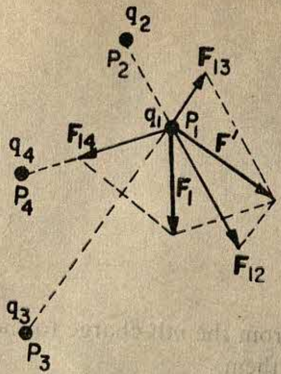


Fig. 1.8 *Total force on charge q_1 due to charges q_2, q_3 and q_4, charge q_1 is repelled by q_2 and q_3 and attracted by q_4.* $\mathbf{F}' = \mathbf{F}_{12} + \mathbf{F}_{13}$, $\mathbf{F}_1 = \mathbf{F}' + \mathbf{F}_{14} = \mathbf{F}_{12} + \mathbf{F}_{13} + \mathbf{F}_{14}$

q_1 due to charges q_2, q_3, q_4, . . . you should proceed as follows: using Coulomb's law (Eq. 1.3a) calculate the force $\mathbf{F}_{12}$ on q_1 due to q_2 alone (ignoring all other charges), then calculate the force $\mathbf{F}_{13}$ on q_1 due to q_3 alone and next calculate the force $\mathbf{F}_{14}$ on q_1 due to q_4, and so on. Finally find the vector sum of all the individual forces using the rules of vector addition you have already learnt in class XI (see Fig. 1.8). Thus the total force on charge q_1 is given by

$$\mathbf{F}_1 = \mathbf{F}_{12} + \mathbf{F}_{13} + \mathbf{F}_{14} + \dots$$

$$= \frac{1}{4\pi\epsilon_0}\left(\frac{q_1q_2}{r_{21}^2}\cdot\hat{\mathbf{r}}_{21} + \frac{q_1q_3}{r_{31}^2}\cdot\hat{\mathbf{r}}_{31} + \frac{q_1q_4}{r_{41}^2}\cdot\hat{\mathbf{r}}_{41} + \dots\right) \qquad (1.4)$$

where $\hat{\mathbf{r}}_{21}$, $\hat{\mathbf{r}}_{31}$, $\hat{\mathbf{r}}_{41}$, . . . are the unit vectors pointing from q_2 to q_1, q_3 to q_1, q_4 to q_1, etc. and r_{21}, r_{31}, r_{41}, . . . are the distances between q_2 and q_1, q_3 and q_1 and q_4 and q_1, etc., respectively. Figure 1.8 shows how the resultant force $\mathbf{F}_1$ is obtained; the forces $\mathbf{F}_{12}$ and $\mathbf{F}_{13}$ are repulsive (i.e. charges q_2 and q_3 are of the same sign as charge q_1) and force $\mathbf{F}_{14}$ is attractive (i.e. charge q_4 has sign opposite to charge q_1).

Similarly the total force on charge q_2 is given by

$$\mathbf{F}_2 = \mathbf{F}_{21} + \mathbf{F}_{23} + \mathbf{F}_{24} + \dots$$

$$= \frac{1}{4\pi\epsilon_0}\left(\frac{q_2q_1}{r_{12}^2}\cdot\hat{\mathbf{r}}_{12} + \frac{q_2q_3}{r_{32}^2}\cdot\hat{\mathbf{r}}_{32} + \frac{q_2q_4}{r_{42}^2}\cdot\hat{\mathbf{r}}_{42} + \dots\right) \qquad (1.5)$$

We can use a more compact notation to represent the force acting on a charge. Suppose we have N point charges q_1, q_2, . . ., q_n labelled $n = 1, 2, \dots, N$ (we could call them 'source' charges) and we wish to find the total force $\mathbf{F}$ on a charge q_0 (which we could call the 'test' charge) due to all the n charges (see Fig. 1.9). The total force can be written compactly, using the summation notation, as

$$\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2 + \mathbf{F}_3 + \dots$$

$$= \sum_{n=1}^{N} \mathbf{F}_n = \frac{1}{4\pi\epsilon_0} \sum_{n=1}^{N} \frac{q_0q_n}{r_n^2}\cdot\hat{\mathbf{r}}_n \qquad (1.6)$$

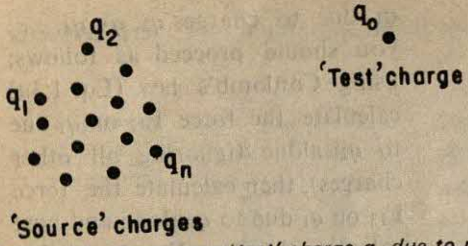


Fig. 1.9 *Force on a 'test' charge q_0 due to n discrete 'source' charges*

where $\hat{\mathbf{r}}_n$ is the unit vector pointing from the *n*th charge to the test charge q_0 and r_n is the distance between them.

Extension to a Continuous Charge Distribution

So far we have discussed situations involving only point charges. We can easily extend Eq. (1.6) to include more general situations where the charge is spread over a region instead of being concentrated at particular points. We have mentioned earlier that charges are quantized and that the minimum magnitude of nonzero charge is that on an electron or a proton. Thus, in this sense, ultimately all charges in nature are quantized and charges less than a minimum value do not exist. But in classical or macroscopic situations we deal with charges which are much greater than the electronic charge. For such situations we may ignore the fact that charge is quantized and imagine the possibility of continuous distributions of charge.

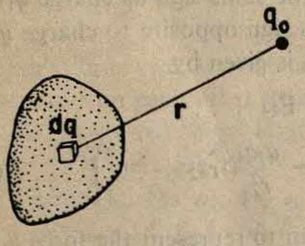


Fig. 1.10 *Calculation of the force on a charge due to a continuous distribution of charge*

Suppose we have a point charge q_0 which is near a region of continuous charge distribution as indicated in Fig. 1.10. To calculate the total force **F** on q_0, we divide the total charge in the region into a very large number of very tiny charges dq distributed over the region and add up (or integrate) vectorially all such contributions. Thus we should write

$$\mathbf{F} = \frac{1}{4\pi\epsilon_0}\int \frac{q_0\,\mathrm{d}q}{r^2}\cdot\hat{\mathbf{r}} = \frac{q_0}{4\pi\epsilon_0}\int \frac{\mathrm{d}q}{r^2}\cdot\hat{\mathbf{r}} \tag{1.7}$$

where $\hat{\mathbf{r}} = \mathbf{r}/r$, is a variable unit vector that points from each elementary charge dq toward the location of the charge q_0.

Different Kinds of Continuous Charge Distributions

There are three kinds of continuous charge distributions that we can think of (see Fig. 1.11): a quantity of charge distributed (i) in a three-dimensional volume or region V of space, (ii) over a two-dimensional surface S in space and (iii) along a one-dimensional curve or line l in space. These are respectively called volume, surface and line distributions.

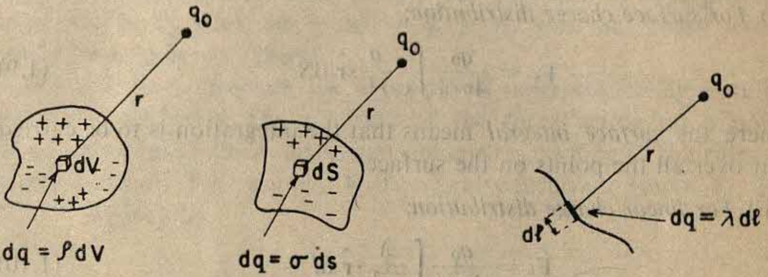


Fig. 1.11 *The three kinds of continuous charge distributions and their densities*

For each kind of charge distribution we define the corresponding *density* of charge. If the charge is continuously distributed inside a volume V (e.g. a sphere or a cube) we define what is called the *volume* density of charge. It is denoted by the symbol ρ and is defined as charge per unit volume, i.e.

$$\rho = \frac{dq}{dV}$$

The SI unit for ρ is coulomb per cubic metre (C m^{-3}). Similarly if the charge is spread continuously over some area S (e.g. a membrane) we define the *surface* density of charge (symbol σ) as the charge per unit area, i.e.

$$\sigma = \frac{dq}{dS}$$

which is expressed in Cm^{-2} in the SI system. And if the charge is uniformly distributed along a line (e.g. a straight line or the circumference of a circle) we define the *linear* density of charge (symbol λ) as the charge per unit length, i.e.

$$\lambda = \frac{dq}{dl}$$

Equation (1.7) of Coulomb's force on a 'test' charge q_0 due to the three kinds of continuous charge distribution can now be written in terms of the density of charge of each distribution.

(i) *For volume charge distribution*:

$$\mathbf{F}_v = \frac{q_0}{4\pi\epsilon_0}\int_V \frac{\rho}{r^2}\cdot\hat{\mathbf{r}}\,\mathrm{d}V \tag{1.8}$$

where we have used what is called the *volume integral* which means that the integration is to be carried out over all the points in volume V.

(ii) *For surface charge distribution*:

$$\mathbf{F}_s = \frac{q_0}{4\pi\epsilon_0}\int_S \frac{\sigma}{r^2}\cdot\hat{\mathbf{r}}\,\mathrm{d}S \tag{1.9}$$

where the *surface integral* means that the integration is to be carried out over all the points on the surface.

(iii) *For linear charge distribution*:

$$\mathbf{F}_l = \frac{q_0}{4\pi\epsilon_0}\int_l \frac{\lambda}{r^2}\cdot\hat{\mathbf{r}}\,\mathrm{d}l \tag{1.10}$$

The *line integral* implies that the integration is to be carried out for all points on the line.

It may be mentioned that the value ρ, σ or λ may be different at different points. In the special cases where density is uniform, ρ, σ and λ have the same value at every point. Remember $\hat{\mathbf{r}} = \mathbf{r}/r$ is a variable unit vector directed from each point in the volume, surface, or line towards the location of the point charge q_0.

Total Electrostatic Force

If the 'test' charge q_0 finds itself in a region which has all kinds of charge distributions such as those shown in Fig. 1.9 (discrete charges) and Fig. 1.11 (three kinds of continuous charge distributions), then the total electrostatic force experienced by the charge q_0 due to all these charge distributions can be obtained by the grand application of the principle of superposition and is given by the vectorial sum of forces given by Eqs (1.6), (1.8), (1.9) and (1.10). Thus

$$\begin{aligned}(\mathbf{F})_{\text{total}} &= \mathbf{F} + \mathbf{F}_V + \mathbf{F}_S + \mathbf{F}_l \\ &= \frac{q_0}{4\pi\epsilon_0}\left[\sum_{n=1}^{N}\frac{q_n}{r_n^2}\cdot\hat{\mathbf{r}}_n + \int_V \frac{\rho}{r^2}\cdot\hat{\mathbf{r}}\,\mathrm{d}V\right. \\ &\quad \left. + \int_S \frac{\sigma}{r^2}\cdot\hat{\mathbf{r}}\,\mathrm{d}S + \int_l \frac{\lambda}{r^2}\cdot\hat{\mathbf{r}}\,\mathrm{d}l\right]\end{aligned} \tag{1.11}$$

The student is advised to refer to Appendix I at the end of the book

Date 20.1.77
Acc. No 9806

where the ideas about line, surface and volume integrals have been explained.

Note: It is easy to understand why the principle of superposition holds for forces that obey Coulomb's law which states that the force on a 'test' charge is proportional to the total 'source' charge. If the force were proportional to the *square* (or any other power different from 1) of the total source charge, the principle of superposition would not hold, since $(q_1 + q_2)^2 \neq q_1^2 + q_2^2$, and the 'cross terms' would have to be considered.

We will now illustrate the principle of superposition by working out a few examples. The resultant of the forces acting on a point charge is equal to their vector sum obtained according to the rules of vector addition. The resultant **F** of several concurrent forces $\mathbf{F}_1$, $\mathbf{F}_2$, $\mathbf{F}_3$, . . . is given by

$$\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2 + \mathbf{F}_3 + \ldots$$

If the forces are coplanar, say in the x-y plane, the x and y components of the resultant force are given by

$$F_x = F_1 \cos \alpha_1 + F_2 \cos \alpha_2 + F_3 \cos \alpha_3 + \ldots$$

and $$F_y = F_1 \sin \alpha_1 + F_2 \sin \alpha_2 + F_3 \sin \alpha_3 + \ldots$$

where α_1, α_2, α_3 . . . respectively are the angles the forces $\mathbf{F}_1$, $\mathbf{F}_2$, $\mathbf{F}_3$, . . . subtend with the positive x-axis.

The magnitude of the resultant vector **F** is given by

$$F = \sqrt{F_x^2 + F_y^2}$$

and its direction is given by

$$\tan \alpha = \frac{F_y}{F_x}$$

where α is the angle subtended by the resultant with the positive x-axis.

Library
Calcutta

EXAMPLE 1.6 Three point charges $+5.0 \times 10^{-6}$ C, -4.0×10^{-6} C and $+4.0 \times 10^{-6}$ C are placed at the vertices P, Q and R of an equilateral triangle of side 10 cm. Find the magnitude and the direction of the total (resultant) force on (a) the charge at P due to charges at Q and R, (b) the charge at Q due to charges at P and R and (c) the charge at R due to charges at P and Q.

Solution: (a) *Force on charge at P due to charges at Q and R*
Referring to Fig. 1.12(a), the magnitude of the attractive force at P

due to the charge at Q is (ignoring the signs of charges, since we are determining the magnitude of the force)

$$F_{PQ} = \frac{1}{4\pi\epsilon_0}\,\frac{q_P q_Q}{r^2}$$

$$= \frac{1}{4\times 3.14\times(8.85\times 10^{-12})}\cdot\frac{(5.0\times 10^{-6})(4.0\times 10^{-6})}{(0.1)^2}$$

$$= 18 \text{ N directed P to Q}$$

Similarly, the repulsive force on P due the charge at R is

$$F_{PR} = 18 \text{ N directed from R to P}$$

The x and y components of the resultant force F on the charge at P are given by [refer to Fig. 1.12(a)]

$$F_x = F_{PQ}\cos\alpha_1 + F_{PR}\cos\alpha_2 \qquad \text{(i)}$$

$$F_y = F_{PQ}\sin\alpha_1 + F_{PR}\sin\alpha_2 \qquad \text{(ii)}$$

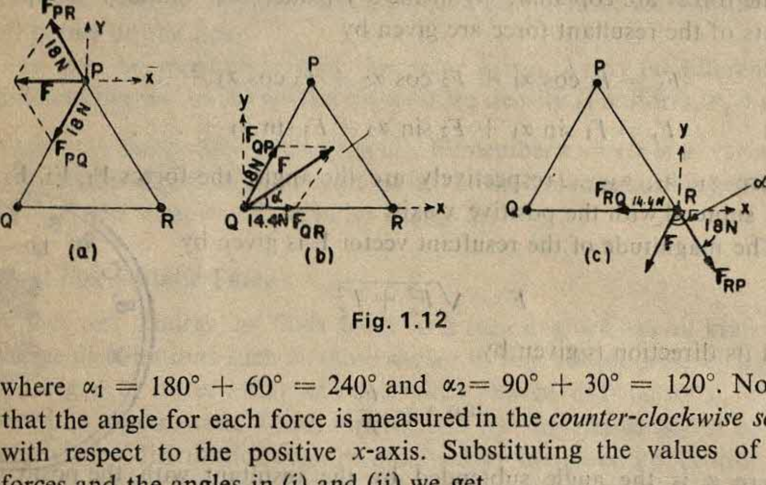


Fig. 1.12

where $\alpha_1 = 180° + 60° = 240°$ and $\alpha_2 = 90° + 30° = 120°$. Notice that the angle for each force is measured in the *counter-clockwise sense* with respect to the positive x-axis. Substituting the values of the forces and the angles in (i) and (ii) we get

$$F_x = 18\times\cos 240° + 18\times\cos 120° = -18 \text{ N}$$

(the $-$ve sign shows that F_x is along the $-$ve x-axis)

$$F_y = 18\times\sin 240° + 18\times\sin 120°$$

$$= 0$$

Therefore, the magnitude of the resultant force on P is

$$F = \sqrt{F_x^2 + F_y^2} = 18 \text{ N}$$

and its direction is along the $-$ve x-axis.

(b) *Force on charge at Q due to charges at P and R* Referring to Fig. 1.12(b), we have

$$F_{QP} = 18 \text{ N directed from Q to P}$$

$$F_{QR} = 14.4 \text{ N directed from Q to R}$$

The x and y components of the resultant force F on the charge at Q are

$$F_x = 14.4 \cos 0^\circ + 18 \cos 60^\circ$$

$$= 23.4 \text{ N}$$

$$F_y = 14.4 \sin 0^\circ + 18 \sin 60^\circ$$

$$= 15.6 \text{ N}$$

$$\therefore \qquad F = \sqrt{F_x^2 + F_y^2} = 28.1 \text{ N}$$

The angle α subtended by the resultant with the +ve x-axis is given by

$$\tan \alpha = \frac{F_y}{F_x} = \frac{15.6}{23.4} = 0.67$$

giving

$$\alpha = 33.7^\circ$$

(c) *Force on charge at R due to charges at P and Q* Referring to Fig. 1.12(c), we have

$$F_{RP} = 18 \text{ N directed from P to R}$$

$$F_{RQ} = 14.4 \text{ N directed from R to Q}$$

$$F_x = 18 \cos 300^\circ + 14.4 \cos 180^\circ = -5.4 \text{ N}$$

$$F_y = 18 \sin 300^\circ + 14.4 \sin 180^\circ = -15.6 \text{ N}$$

$$\therefore \qquad F = \sqrt{F_x^2 + F_y^2} = 16.5 \text{ N}$$

$$\tan \alpha = \frac{F_y}{F_x} = \frac{-15.6}{-5.4} = 2.9$$

giving

$$\alpha = 70.9^\circ \simeq 71^\circ$$

Example 1.7 Two negative charges each of magnitude 8×10^{-8} C and a positive charge q are placed along a straight line. At what position and for what value of q will the system be in equilibrium. Is the equilibrium stable or unstable?

Solution: Let the positive charge q be placed at R at distances r_1 and r_2 respectively from the two negative charges ($q_0 = 8 \times 10^{-8}$ C) at P and Q as shown in Fig. 1.13.

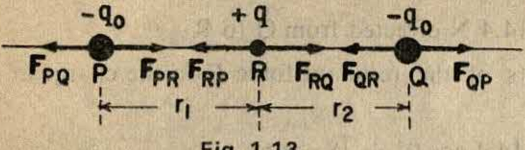


Fig. 1.13

Equilibrium of q For charge q to be in equilibrium, it must lie between P and Q and the forces on it exerted by the charge q_0 at P and Q must be co-linear, equal and opposite, i.e.

$$|F_{RP}| = |F_{RQ}|$$

or

$$\frac{1}{4\pi\epsilon_0} \cdot \frac{qq_0}{r_1^2} = \frac{1}{4\pi\epsilon_0} \cdot \frac{qq_0}{r_2^2}$$

giving $r_1 = r_2 = r$ (say). Hence charge q should be equidistant from P and Q.

Equilibrium of q_0 at P and Q For the system to be in equilibrium, the charge q_0 at P and Q must also be in equilibrium, i.e.

$$|F_{PR}| = |F_{PQ}| \quad \text{and} \quad |F_{QR}| = |F_{QP}|$$

Both these conditions give $q = q_0/4$. Thus for q_0 at P and Q to be in equilibrium, the value of charge

$$q = \frac{q_0}{4} = 2 \times 10^{-8} \text{ C}$$

Stability of equilibrium A slight displacement of q towards P increases the magnitude of F_{RP} and decreases the magnitude of F_{RQ}. Consequently, the displacement of q will increase still further towards P. The three charges are no longer in equilibrium. Hence the original equilibrium is *unstable* for displacement along the line joining the three charges.

For a displacement of q along a direction perpendicular to the line joining P and Q, the resultant of the two forces of attraction F_{RP} and F_{RQ} acting on charge q will bring it back to its original position. The equilibrium is stable for displacement in the direction perpendicular to the line joining the three charges.

EXAMPLE 1.8 Two small identical balls P and Q, each of mass 0.20 g, carry identical charges and are suspended by two threads of equal lengths. At equilibrium, they position themselves as shown in Fig. 1.14. Find the charge on each ball. Given $g = 9.8$ ms^{-2}.

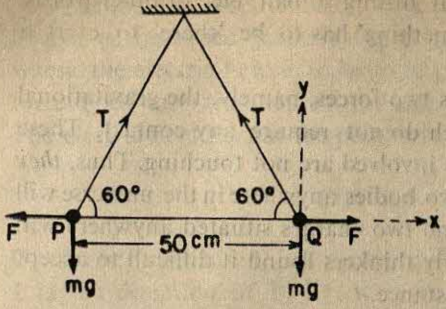


Fig. 1.14

Solution: Let us consider the ball, say, Q. There are three forces acting on it: (i) tension T in the thread, (ii) the force mg due to gravity acting vertically downwards and (iii) the force F due to Coulomb's repulsion in the $+x$ direction.

For equilibrium, the sum of the x and y components of these forces must be zero, i.e.

$$T \cos 60^\circ - F = 0 \qquad \text{(i)}$$

and

$$T \sin 60^\circ - mg = 0 \qquad \text{(ii)}$$

From (ii) we have

$$T = \frac{mg}{\sin 60^\circ} = \frac{(0.20 \times 10^{-3}) \times 9.8}{0.866} = 2.26 \times 10^{-3} \text{ N}$$

Substituting this value of T in (i) we get

$$F = T \cos 60^\circ = (2.26 \times 10^{-3}) \times 0.5 = 1.13 \times 10^{-3} \text{ N}$$

Let the charge on each ball be q. The Coulomb's force of repulsion between them is $F = 1.13 \times 10^{-3}$ N. Hence

$$1.13 \times 10^{-3} = \frac{1}{4\pi\epsilon_0} \cdot \frac{q^2}{r^2}$$

Now $r = 50$ cm $= 0.50$ m and $\frac{1}{4\pi\epsilon_0} = 9 \times 10^9 \text{ Nm}^2 \text{ C}^{-2}$. Substituting these values, we get

$$q = 1.77 \times 10^{-7} \simeq 1.8 \times 10^{-7} \text{ C}$$

1.11 THE ELECTRIC FIELD: A USEFUL CONCEPT

Most of the forces that we come across in daily life are the so-called 'contact' forces. There is a contact between the object that exerts the force and one that feels it. For example, a man pushing a table, a hammer driving a nail or a bat hitting a ball, etc. In other words, when an object feels a force, 'something' has to be 'there' to exert a force.

However, we have come across two forces, namely, the gravitational force and the electric force, which do not require any contact. These forces exist even when the bodies involved are not touching. Thus, *they act at a distance*. For example, two bodies anywhere in the universe will experience gravitational force and two charges situated anywhere will experience electric force. The early thinkers found it difficult to accept the idea of a force acting at a distance.

The difficulty can be reduced if we introduce the idea of a field. We imagine that an electric charge gives rise to an electric field in the entire space around it. Therefore, the force that one charge exerts on another can be described as the force that one charge experiences due to the electric field set up by the other. The electric field may be regarded as a modification of the space brought about by the presence of a charge. It is a space property. Thus, when a charge feels a force, 'something' must be 'there' to exert the force. This 'something' is the electric field of some other charge (or a collection of charges). The usefulness of this concept lies in the fact that even if we could not see the charged body (it may be behind a wall) we could still feel its existence because of its electric field.

To find out whether there is any electric field at any point in space, we place a test charge at that point and observe whether it experiences any force. If the test charge does not experience any force, there is no electric field at that point. If the test charge experiences a force, there, must exist an electric field at that point (due to some other charge somewhere).

Definition of Electric Field

To give an operational definition of electric field, let us consider the analogy of gravitational force. With every point in space near the earth we can associate a *gravitational field* $\mathbf{g}(\mathbf{r})$. This is the gravitational acceleration that a test body placed at that point would experience. If m is the mass of the body placed at $\mathbf{r}$, $\mathbf{F}$ is the gravitational force acting on it, the gravitational field $\mathbf{g}(\mathbf{r})$ is given by

$$\mathbf{g}(\mathbf{r}) = \frac{\mathbf{F}}{m} \tag{1.12}$$

Similarly, with every point in space surrounding a charge, we can associate an *electric field*. In the same way we speak of a *magnetic field* at every point in the space surrounding a magnet. The electric and magnetic fields are central concepts in classical electromagnetism.

To define the electric field, we place a test body carrying a test charge q_0 (assumed positive for convenience) at the point in space where the electric field is to be found, and we measure the Coulomb's force **F** (if any) that acts on the test charge due to the source charge. The electric field **E**(**r**) at the point **r** is space is given by

$$\mathbf{E}(\mathbf{r}) = \frac{\mathbf{F}}{q_0} \tag{1.13}$$

E is a vector because **F** is a vector, q_0 being a scalar. The direction of **E** is the direction of **F**, i.e. it is in the direction in which a resting positive charge placed at the point would tend to move. The gravitational field **g** defined by Eq. (1.12) is also a vector field.

Notice from Eqs (1.12) and (1.13) that the definition of electric field **E** is similar to that of gravitational field **g**, except that the charge of the test body, rather than its mass, is the quantity of interest. The symbol **E**(**r**) is used to indicate that **E** depends upon the position **r** of the test charge.

Although the unit of **g** is usually written as ms^{-2}, we could also write it as newtons per kilogram (N kg^{-1}). The SI unit of electric field **E** is similarly *newtons per coulomb* (NC^{-1}). Thus both **g** and **E** are expressed as a force divided by the relevant property (mass or charge) of the test body. Remember that **E** is the electric field 'created' by a 'source' charge which could be a discrete (Fig. 1.9) and/or a continuous distribution (Fig. 1.11).

A critical reader will immediately find a flaw in our definition of electric field. He will ask: will the test charge q_0 not have its own electric field? Will it, therefore, not modify the electric field of the source charge? The answer is that it certainly will. The charge q_0 would definitely exert a Coulomb force on every charge constituting the source which might make the latter move and change their positions. To minimise and ultimately remove this effect, the electric field **E**(**r**) should, strictly speaking, be defined as

$$\mathbf{E}(\mathbf{r}) = \underset{q_0 \to 0}{\text{Limit}} \ \frac{\mathbf{F}}{q_0} \tag{1.14}$$

where the test charge q_0 is so small that its effect on the electric field of the source charge can be neglected. The idea of taking the limit $q_0 \to 0$ is that the electric field **E**(**r**) is more and more exactly defined as the value of q_0 is made smaller and smaller, and finally vanishes. Notice that q_0 cannot be exactly equal to zero for then it would not experience any Coulomb force and there would be no way of measuring the electric field.

The idea of taking the limit $q_0 \rightarrow 0$ conflicts with the fact that in nature there is no smaller positive charge than that on a proton ($= 1.6 \times 10^{-19}$ C). However, as stated earlier, while dealing with macroscopic (classical) situations, we can safely ignore the fact that charge is quantized and that a charge smaller than a minimum value cannot exist. We imagine that the charge can approach zero in a continuous way.

Physical Significance and Use of Electric Field

In principle the electric field $\mathbf{E}(\mathbf{r})$ can be calculated if all the source charges and their locations are known. The idea of the electric field, defined all over the space around the source charge (say Q), is just a convenient way of knowing in advance what force would act on a charge q_0 if it were placed at different positions in space because

$$\mathbf{F}(r) = q_0\mathbf{E}(\mathbf{r})$$

Thus we can visualize the electric field playing an intermediary role in the forces between charges. The source charge Q sets up an electric field $\mathbf{E}(\mathbf{r})$ at all points $\mathbf{r}$ in the space around it, the value of this field at any point then determines what force a test charge q_0 at that point will experience, i.e. we think in terms of

$$\text{Charge} \rightleftharpoons \text{Electric field} \rightleftharpoons \text{Charge}$$

It is in this sense that the concept of electric field is useful.

In Chapter 5 we will use a similar method to define magnetic field. We have so far dealt with *stationary* charges so that the field due to a charge is a static field whose value at each point does not change with time. If the only problem in electromagnetism was that of forces between stationary charges, we would have never used the field concept at all. When we discuss electromagnetic induction (Chapter 7) and electromagnetic waves (Chapter 9) we will have to work with more general vector fields which are not static, i.e. their values at each point of space could change with time.

Let us consider an example of a time-varying field. Suppose that the source charge Q suddenly moves towards the test charge q_0. How quickly will the charge q_0 'learn' that the charge Q has moved? Electromagnetic theory predicts that q_0 'learns' about the motion of Q by a field disturbance that starts from charge Q and travels with the speed of light. Experiments also show that the information about Q's motion is not *instantaneously* communicated to q_0 (see Chapter 9). Accelerating electrons in the antenna of a radio transmitter influence electrons in a distant receiving antenna only after a time x/c where x is the distance between the two antennas and c is the speed of light.

EXAMPLE 1.9 Figure 1.15 shows a point charge $q_1 = +1.0 \times 10^{-8}$ C at a distance $l = 10$ cm from another point charge $q_2 = +2.0 \times 10^{-8}$ C. At what point on the line joining the two charges is the electric field zero?

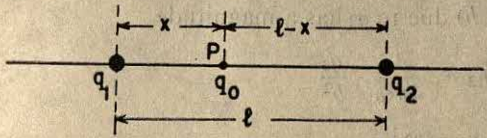


Fig. 1.15

Solution: The electric field will be zero at a point where the forces exerted by q_1 and q_2 on a test charge q_0 are equal and opposite. This point P must, therefore lie between the charges. If F_1 is the force due to q_1 and F_2 that due to q_2, we must have

$$F_1 = F_2$$

or

$$\frac{1}{4\pi\epsilon_0} \cdot \frac{q_0 q_1}{x^2} = \frac{1}{4\pi\epsilon_0} \cdot \frac{q_0 q_2}{(l-x)^2}$$

where x is the distance of P from q_1 and $l = 10$ cm $= 0.1$ m. Thus we have

$$\frac{q_1}{x^2} = \frac{q_2}{(l-x)^2}$$

Putting the values of q_1, q_2 and l we get

$$x^2 + 0.2x - 0.01 = 0$$

The two roots of this quadratic equation are

$$x = 0.041 \text{ m and } -0.241 \text{ m}$$

The negative root is inadmissible because it corresponds to a point to the left of q_2 at a distance 0.241 m from it where the electric field cannot be zero. Hence the electric field is zero at a point P between the two charges at a distance of 0.041 m = 4.1 cm from charge q_1.

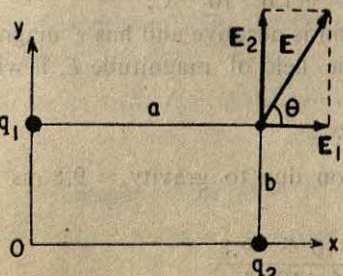


Fig. 1.16

EXAMPLE 1.10 Two point charges $q_1(= +2.0\,\mu C)$ and $q_2(= +1.0\,\mu C)$ are placed at distances $a = 2.0$ m and $b = 1.0$ m from the origin on the x and y axes as shown in Fig. 1.16. Find the electric field **E** at point (a, b).

Solution: The electric field $\mathbf{E}_1$ at (a), (b) due to q_1 has a magnitude

$$E_1 = \frac{1}{4\pi\epsilon_0} \cdot \frac{q_1}{a^2}$$

and is directed along the $+x$ – axis.

The electric field $\mathbf{E}_2$ at (a, b) due to q_2 has a magnitude

$$E_2 = \frac{1}{4\pi\epsilon_0} \cdot \frac{q_2}{b^2}$$

and is directed along the $+y$ – axis.

Since the two vectors $\mathbf{E}_1$ and $\mathbf{E}_2$ are at right angles to each other, the magnitude of the resultant electric field $\mathbf{E}$ is given by

$$E = \sqrt{E_1^2 + E_2^2}$$

$$= \frac{1}{4\pi\epsilon_0}\left[\left(\frac{q_1}{a^2}\right)^2 + \left(\frac{q_2}{b^2}\right)^2\right]^{1/2}$$

Substituting the values of q_1, q_2, a, b and ϵ_0 we get

$$E = 7.8 \times 10^3 \text{ NC}^{-1}$$

The angle θ subtended by the resultant field $\mathbf{E}$ with the x-axis is given by

$$\tan\theta = \frac{E_2}{E_1} = \frac{q_2}{q_1} \cdot \frac{a^2}{b^2} = \frac{1}{2} \times \left(\frac{2}{1}\right)^2 = 2.0$$

giving $\theta = 63.4^\circ$.

EXAMPLE 1.11 (a) An electron is placed in a uniform electric field near the surface of the earth. What must be the magnitude and the direction of the electric field so that the electron experiences an electrical force equal to its weight? (b) What will happen to a proton placed in this field? Given, mass of an electron $(m_e) = 9.1 \times 10^{-31}$ kg, mass of a proton $(m_p) = 1.67 \times 10^{-27}$ kg, charge on an electron $= -1.6 \times 10^{-19}$ C, charge on a proton $= +1.6 \times 10^{-19}$ C.

Solution: (a) The charge on an electron is negative and has a magnitude $e = 1.6 \times 10^{-19}$ C. In an electric field of magnitude E, it will experience an electric force F given by

$$F = eE$$

but $F = m_e g$ where g is the acceleration due to gravity $= 9.8 \text{ ms}^{-2}$. Hence E is given by

$$E = \frac{F}{e} = \frac{m_e g}{e} = \frac{9.1 \times 10^{-31} \times 9.8}{1.6 \times 10^{-19}}$$

$$= 5.6 \times 10^{-11} \text{ NC}^{-1}$$

Since the charge on an electron is negative and the electric force must be vertically upwards (so that it can balance its weight acting vertically downwards), the direction of electric field must be vertically downwards. By definition, the direction of **E** is the direction in which a positive charge would move when placed in that field.

(b) When a proton which has a positive charge is placed in this electric field, it will experience a downward electrical force given by

$$\begin{aligned}(F_{\text{elec}}). &= eE\\ &= 1.6\times10^{-19}\times5.6\times10^{-11}\\ &= 8.96\times10^{-30}\ \text{N}\end{aligned}$$

It addition, it will experience another downward force due to the gravity of the earth

$$\begin{aligned}(F)_{\text{grav.}} = m_{\text{p}}\times g &= 1.67\times10^{-27}\times9.8\\ &= 1.64\times10^{-26}\ \text{N}\end{aligned}$$

Thus the electrical force on the proton is negligible compared to the gravitational force. Hence the proton will fall almost freely at an acceleration of 9.8 ms^{-2}. The downward acceleration due to the electric force is

$$\begin{aligned}(a)_{\text{elec.}} = \frac{(F)_{\text{elec}}}{m_{\text{p}}} &= \frac{8.96\times10^{-30}}{1.67\times10^{-27}}\\ &= 5.4\times10^{-3}\ \text{ms}^{-2}\end{aligned}$$

Notice that the acceleration of a charged body in an electric field depends upon its mass, but its acceleration due to gravity does not.

EXAMPLE 1.12 An oil drop having 12 excess electrons is held stationary under a uniform electric field of $2.55\times10^{4}\ \text{Vm}^{-1}$ in Millikan's oil drop experiment. The density of the oil is $1.26\ \text{gcm}^{-3}$. Estimate the radius of the drop. Given $g = 9.81\ \text{ms}^{-2}$, $e = 1.60\times10^{-19}$ C.

Solution: Charge on the oil drop $(q_0) = 12e = 12\times1.60\times10^{-19}$

$$= 1.92\times10^{-18}\ \text{C}$$

Density of oil $(\rho) = 1.26\ \text{g cm}^{-3} = 1.26\times10^{3}\ \text{kgm}^{-3}$

Mass of the oil drop (m) = volume × density

$$= \frac{4\pi}{3}r^3\rho$$

where r is the radius of the drop in metres. The electric field is

$$E = 2.55 \times 10^4 \text{ Vm}^{-1} = 2.55 \times 10^4 \text{ NC}^{-1}$$

$$\left(\text{volt} = \frac{\text{joule}}{\text{coulomb}} = \frac{\text{Nm}}{\text{C}}, \quad \therefore \quad \text{NC}^{-1} = \text{Vm}^{-1}\right)$$

The field is directed vertically downwards so that the negatively charged drop experiences an upward force due to the electric field. This force is

$$F_e = q_0 E$$

The downward gravitational force on the drop is

$$F_g = mg$$

The drop will be stationary if

$$F_e = F_g$$

or
$$q_0 E = mg$$

or
$$q_0 E = \frac{4\pi}{3} r^3 \rho g$$

or
$$r = \left(\frac{3 q_0 E}{4\pi \rho g}\right)^{1/3}$$

Substituting the values of q_0, E and ρ we get

$$r = 9.81 \times 10^{-7} \text{ m} = 9.81 \times 10^{-4} \text{ mm}$$

Note: We have ignored the buoyant force on the drop compared to F_e or F_g. Since the volume of the drop is very small and the density of air ($= 1.29 \text{ kgm}^{-3}$) is negligible compared to that of oil, the buoyant force will be too small to influence the drop.

EXAMPLE 1.13 In a cathode ray oscilloscope, an electron accelerated through 25 kV enters a deflecting plate system as shown in Fig. 1.17.

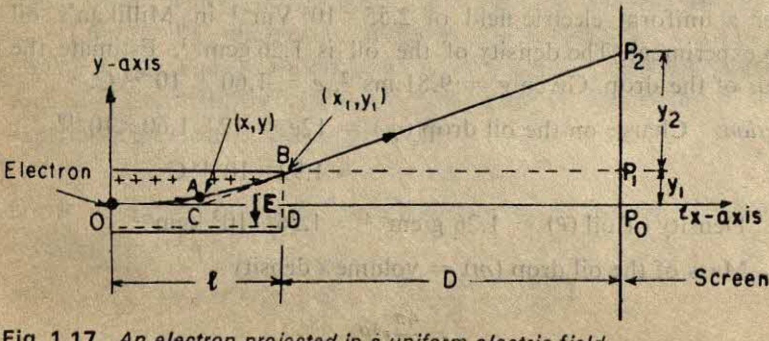


Fig. 1.17 *An electron projected in a uniform electric field*

The magnitude of the deflecting field (E) is 20 kVm^{-1}. The length of the plate (l) is 5.0 cm and the screen is 18 cm (D) away from the nearer end of the plate. (a) What is the nature of electron's trajectory in the region between the plates? (b) At what point does the electron hit the fluorescent screen?

Solution: (a) Let an electron of mass m and charge e be projected with a velocity v_x along the x-axis between two plates where a uniform electric field exists pointing downward as shown in Fig. 1.17. The motion of the electron is similar to that of a projectile fired horizontally in the earth's gravitational field, except that it moves upwards against the electric field. The horizontal (x) and vertical (y) motions are given by

$$x = v_x t$$

and

$$y = \frac{1}{2}at^2 = \frac{eEt^2}{2m}$$

where t is the time taken by the electron to reach a point A and a is its acceleration (directed upwards) at this point. Eliminating t yields

$$y = \frac{eE}{2mv_x^2}\cdot x^2 \tag{i}$$

This is the equation of a parabola. Hence the trajectory of the electron (OAB) within the plates is a parabola.

(b) Potential difference between plates (V_0) = 25 kV

$$= 25 \times 10^3 \text{ V}$$

Kinetic energy of the electron at entry ($\frac{1}{2}mv_x^2$)

$$= eV_0 \tag{ii}$$

When the electron emerges from the end B of the plates, it has suffered a deflection y_1 given by [see (i)]

$$y_1 = \frac{eE}{2mv_x^2}\cdot x_1^2$$

$$= \frac{eEl^2}{2mv_x^2} \qquad (\because\ x_1 = l)$$

Using Eq. (ii) we have

$$y_1 = \frac{eEl^2}{4eV_0} = \frac{El^2}{4V_0} \tag{iii}$$

When the electron emerges from the plates (point B), it travels essentially in a straight line tangent to the parabola at the exit point B. The effect of gravity is negligible since the mass of an electron is very small. Let it strike the screen at point P_2. P_2B produced will meet the x-axis at point C almost mid-way between the ends of the plates, i.e. $CD = l/2$. The total deflection y of the electron consists of two parts (see Fig. 1.17).

(i) deflection y_1 within plates and (ii) deflection y_2 outside the plates

$$y = y_1 + y_2$$

Now the triangles P_2BP_1 and BCD are similar. Hence

$$\frac{P_2P_0}{P_0C} = \frac{BD}{CD}$$

or
$$\frac{y}{\left(D + \frac{l}{2}\right)} = \frac{y_1}{l/2}$$

or
$$y = \frac{2y_1}{l}\left(D + \frac{l}{2}\right) \qquad \text{(iv)}$$

Using (iii) in (iv) we get

$$y = \frac{El}{2V_0}\left(D + \frac{l}{2}\right) \qquad \text{(v)}$$

Now $E = 20\ \text{kVm}^{-1} = 20 \times 10^3\ \text{Vm}^{-1}$, $V_0 = 25 \times 10^3$ V

$l = 5.0$ cm $= 5.0 \times 10^{-2}$ m, $\quad D = 18$ cm $= 18 \times 10^{-2}$ m.

Substituting these values in Eq. (v) yields

$$y = 4.1\ \text{mm}$$

Example 1.14 A conducting rod of negligible thickness and length L carried a total charge q which is uniformly distributed on it. Find the electric field at a point located on the axis of the rod at a distance a from the nearer end.

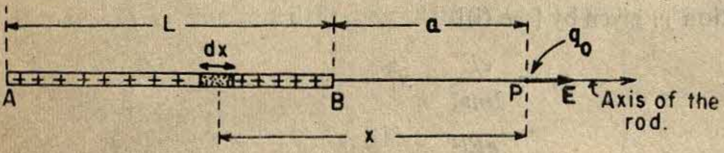


Fig. 1.18

Solution: Let us consider a small element dx of the rod located at a distance x from point P where the electric field is to be determined. Since the charge q is distributed uniformly on the rod of length L, the charge of the element of length dx is

$$dq = \frac{q}{L}\,dx$$

Hence the force on a test charge q_0 at P due to dq is

$$dF = \frac{1}{4\pi\epsilon_0} \cdot \frac{dq q_0}{x^2} = \frac{1}{4\pi\epsilon_0} \cdot \frac{qq_0}{L} \cdot \frac{dx}{x^2}$$

The magnitude of the total force on q_0 due to the entire rod is given by

$$F = \int dF = \int_a^{(L+a)} \frac{1}{4\pi\epsilon_0} \cdot \frac{qq_0}{L} \cdot \frac{dx}{x^2}$$

$$= \frac{qq_0}{4\pi\epsilon_0 L} \int_a^{(L+a)} \frac{dx}{x^2}$$

$$= \frac{qq_0}{4\pi\epsilon_0 a(L+a)}$$

Hence the magnitude of the electric field at P is

$$E = \frac{F}{q_0} = \frac{q}{4\pi\epsilon_0 a(L+a)}$$

If the charge q is positive, the direction of the field will be from left to right. If $L \ll a$, the expression for E reduces to

$$E = \frac{q}{4\pi\epsilon_0 a^2}$$

which varies inversely as the square of the distance a. This is expected since when $L \ll a$, the rod can be regarded as a point charge.

EXAMPLE 1.15 A loop of wire of negligible thickness has a radius r. It carries a charge q distributed uniformly over the loop. Find the electric field at a point P located at a distance x from the centre along the line perpendicular to the plane containing the loop and passing through its centre.

Solution: Let us suppose that the loop of the wire is placed such that the plane of the loop is perpendicular to the x-axis as shown in Fig. 1.19. Consider a small element (of length dl) of the loop. Since the

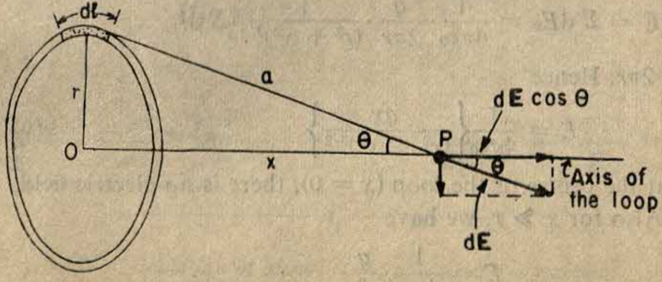


Fig. 1.19

total charge q is distributed uniformly, the charge dq on the element dl is

$$\mathrm{d}q = \frac{q}{2\pi r} \cdot \mathrm{d}l \qquad \text{(i)}$$

The magnitude dE of the electric field d**E** at point P due to charge dq is (see Fig. 1.19)

$$\mathrm{d}E = \frac{1}{4\pi\epsilon_0} \cdot \frac{\mathrm{d}q}{a^2} \qquad \text{(ii)}$$

where $a = (x^2 + r^2)^{1/2}$. Using Eq. (i) in (ii) we have

$$\mathrm{d}E = \frac{1}{4\pi\epsilon_0} \cdot \frac{q}{2\pi_r} \cdot \frac{\mathrm{d}l}{(r^2 + x^2)} \qquad \text{(iii)}$$

The total electric field **E** at P is found by adding, vectorially, the d**E** vectors for all the elements that make up the loop. At first sight, it is a difficult task because the vectors do not lie in the same plane. However, the symmetry about the axis of the loop makes our job very easy indeed. Only the components of d**E** that lie along the axis contribute. The components that are perpendicular to this axis add up to zero. The axial components are all collinear, so the vector addition becomes a scalar addition of parallel axial components.

The axial, i.e. the x-component of d**E**, from Fig. 1.19, has a magnitude $\mathrm{d}E_x = \mathrm{d}E \cos\theta$, where $\cos\theta$ is given by

$$\cos\theta = \frac{x}{(r^2 + x^2)^{1/2}}$$

Thus from Eq (iii) we have

$$\begin{aligned} \mathrm{d}E_x &= \mathrm{d}E \cos\theta \\ &= \frac{1}{4\pi\epsilon_0} \cdot \frac{qx}{2\pi r} \cdot \frac{\mathrm{d}l}{(r^2 + x^2)^{3/2}} \qquad \text{(iv)} \end{aligned}$$

To add the various contributions, we need to add only dl's. Thus the addition yields

$$E = \Sigma\, \mathrm{d}E_x = \frac{1}{4\pi\epsilon_0} \cdot \frac{q}{2\pi r} \cdot \frac{1}{(r^2 + x^2)^{3/2}} \cdot \Sigma\, \mathrm{d}l$$

Now $\Sigma\, \mathrm{d}l = 2\pi r$. Hence

$$E = \frac{1}{4\pi\epsilon_0} \left\{ \frac{qx}{(r^2 + x^2)^{3/2}} \right\} \qquad \text{(v)}$$

Notice that at the centre of the loop ($x = 0$), there is no electric field, as expected. Also for $x \gg r$, we have

$$E = \frac{1}{4\pi\epsilon_0} \cdot \frac{q}{x^2}$$

which implies that for points on the axis located at distances much larger than the dimensions of the loop, it can be regarded as a point charge.

1.12 THE ELECTRIC DIPOLE

A pair of equal and opposite point charges separated by a vector distance **a** *is called an electric dipole or simply a dipole.* The vector **a** is drawn from the negative to the positive charge and is along the *axis* of the dipole. The axis of the dipole is the line joining the two point charges. Calculation of the electric field of a dipole is an easy problem as we show below; it does not involve integration because only point charges are involved. Why is this particular configuration of two charges so important as to merit special consideration? The reason is that dipoles are a common occurrence in nature. A molecule made up of a positive and a negative ion is an example of an electric dipole. Also, a complicated array of charges can be described and analysed in terms of dipoles.

We will now calculate the electric field of the dipole at a point (i) on its axis and (ii) on the perpendicular bisector of the line joining the charges. For convenience, let us place the dipole along the x-axis at the origin of the coordinate system as shown in Fig. 1.20. By symmetry, the results along the y and z axes will be identical, so we restrict our discussion to the x and y axes.

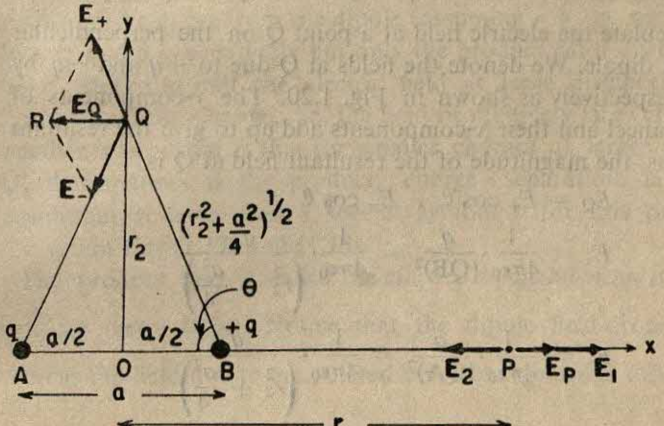


Fig. 1.20 *Calculation of electric field of a dipole of moment* $\mathbf{p}=qa$

Electric Field at a Point on the Axis of a Dipole

We first calculate electric field at a point P at a distance r_1 from the centre of the dipole. The electric $\mathbf{E}_1$ at P due to charge $+q$ has a magnitude

$$E_1 = \frac{1}{4\pi\epsilon_0}\cdot\frac{q}{\left(r_1 - \frac{a}{2}\right)^2}$$

and is directed along BP.

The electric field $\mathbf{E}_2$ at P due to charge $-q$ has a magnitude

$$E_2 = \frac{1}{4\pi\epsilon_0}\cdot\frac{q}{\left(r_1 + \frac{a}{2}\right)^2}$$

and is directed along PA.

The resultant electric field $\mathbf{E}_p$ at P due to two charges at A and B is

$$E_p = E_1 - E_2$$

$$= \frac{1}{4\pi\epsilon_0}\left[\frac{q}{\left(r_1 - \frac{a}{2}\right)^2} - \frac{q}{\left(r_1 + \frac{a}{2}\right)^2}\right]$$

or

$$E_p = \frac{(qa)}{4\pi\epsilon_0}\cdot\frac{2r_1}{\left(r_1^2 - \frac{a^2}{4}\right)^2} \tag{1.15}$$

which is directed along BP.

Electric Field at a Point on the Perpendicular Bisector of a Dipole

Next, we calculate the electric field at a point Q on the perpendicular bisector of a dipole. We denote the fields at Q due to $+q$ and $-q$ by $\mathbf{E}_+$ and $\mathbf{E}_-$ respectively as shown in Fig. 1.20. The y-components of these fields cancel and their x-components add up to give the resultant field $\mathbf{E}_Q$. Thus, the magnitude of the resultant field at Q is

$$E_Q = E_+ \cos\theta + E_- \cos\theta$$

where

$$E_+ = \frac{1}{4\pi\epsilon_0}\cdot\frac{q}{(\mathrm{QB})^2} = \frac{1}{4\pi\epsilon_0}\cdot\frac{q}{\left(r_2^2 + \frac{a^2}{4}\right)}$$

$$E_- = \frac{1}{4\pi\epsilon_0}\cdot\frac{q}{(\mathrm{QA})^2} = \frac{1}{4\pi\epsilon_0}\cdot\frac{q}{\left(r_2^2 + \frac{a^2}{4}\right)}$$

and

$$\cos\theta = \frac{a/2}{\left(r_2^2 + \frac{a^2}{4}\right)^{1/2}}$$

The substitution of these expressions yields

$$E_Q = \frac{1}{4\pi\epsilon_0} \cdot \frac{(qa)}{\left(r_2^2 + \frac{a^2}{4}\right)^{3/2}} \tag{1.16}$$

which is directed along QR.

Electric Dipole Moment

The expressions (1.15) and (1.16) are simplified if the points of interest are far from the dipole compared with the dipole dimension a. When $r_1 \gg a$ and $r_2 \gg a$, the term $a^2/4$ may be dropped in the denominators and we get much simpler expressions.

$$E_P = \frac{1}{4\pi\epsilon_0} \cdot \frac{2(qa)}{r_1^3} = \frac{1}{4\pi\epsilon_0} \cdot \frac{2p}{r_1^3} \tag{1.17}$$

$$E_Q = \frac{1}{4\pi\epsilon_0} \cdot \frac{(qa)}{r_2^3} = \frac{1}{4\pi\epsilon_0} \cdot \frac{p}{r_2^3} \tag{1.18}$$

where $\quad p = qa.$ (1.19)

This approximation is of great practical importance. In atomic, molecular and solid state physics, we are often interested in fields of dipoles at distances very large compared with the dimensions of atomic dipoles. For such situations, Eqs (1.17) and (1.18) hold.

An essential feature of the dipole field is that the magnitude of the charge q and the separation a appear only as product. We have used the symbol p for this product (qa). This means that if we measure **E** at various distances from the dipole (assuming $r \gg a$), we can never deduce q and a separately but only the product qa; if q were doubled and a reduced to half, the electric field at large distances from the dipole would not change, i.e. the field of relatively large charges close together is the same as that for smaller charges at larger separation. All that matters is the product: charge × separation. It is for this reason that we have used a special symbol p for this product (i.e. $p = qa$) in Eqs (1.17) and (1.18).

The product (qa) is called the *electric dipole moment*; its SI unit is coulomb metre (Cm). Notice that the dipole field drops off as $\frac{1}{r^3}$, whereas the field due to an isolated point charge drops off as $1/r^2$.

A Dipole in an Electric Field

The electric dipole moment can be regarded as a vector **p** whose magnitude p is the product of the magnitude of either charge q and the distance a between the charges. The direction of $\mathbf{p}$ is taken to be from the *negative* to the *positive* charge, i.e. the separation $\mathbf{a}$ is a vector from the *negative* to the *positive* charge. Thus

$$\mathbf{p} = q\mathbf{a} \tag{1.20}$$

Suppose a certain electric field has been produced by some external sources and a dipole of dipole moment $\mathbf{p}$ is placed in that field. Let us consider the electric field to be uniform, i.e. it has the same constant magnitude and direction at all points. For a general orientation of the dipole, it will make a certain angle θ with the electric field $\mathbf{E}$. This is the angle between the dipole moment $\mathbf{p}$ and E as shown in Fig. 1.21.

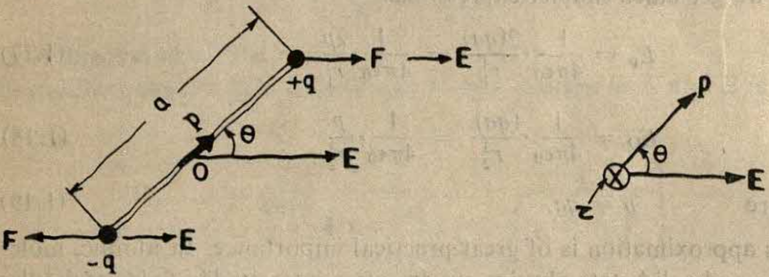


Fig. 1.21 *An electric dipole in a uniform electric field*

To find out what forces act on the dipole, we can picture a dipole configuration as a thin, rigid insulating rod of length a with charges $+q$ and $-q$ stuck at its ends, as shown in the diagram. The electric field E exerts a force $\mathbf{F} = q\mathbf{E}$ on charge $+q$ and a force $\mathbf{F} = -q\mathbf{E}$ on charge $-q$. These two forces at the end are equal and opposite and add up to zero. So the *net force on a dipole in a uniform electric field is zero.*

Torque on a dipole in a uniform field

A very important property of an electric dipole is the torque exerted on it by a uniform electric field. Although the two forces acting at the ends cancel each other, *they act at different points* and hence give rise to a *torque* on the dipole. Since the net force on the dipole is zero, it cannot have any translational motion. We can, therefore, assume that the centre (of mass) of the dipole remains at rest. The torque will rotate the dipole so as to reduce the angle θ to zero (see Fig. 1.21), i.e. $\mathbf{p}$ will tend to align with $\mathbf{E}$. Thus *the torque on a dipole due to an external electric field tends to align the dipole moment parallel to the electric field.*

The magnitude τ of the torque $\boldsymbol{\tau}$ with respect to the centre of the dipole is the sum of the two forces times their moment arms:

$$\tau = F\left(\frac{a}{2}\sin\theta\right) + F\left(\frac{a}{2}\sin\theta\right)$$

$$= Fa\sin\theta = qaE\sin\theta$$

or

$$\tau = pE\sin\theta \tag{1.21}$$

Equation (1.21) is incomplete because it does not state explicitly the sense of the torque for all relative orientations of **p** and **E**. It is more useful, therefore, to cast this equation into the vector form which is completely explicit. The definition of vector product of two vectors can be used to write Eq. (1.21) in vector form as

$$\boldsymbol{\tau} = \mathbf{p} \times \mathbf{E} \tag{1.22}$$

The magnitude of the torque vector $\boldsymbol{\tau}$ is $pE \sin\theta$ and its direction is given by the right-hand screw rule which tells us that the torque vector points along the axis of rotation (which would be produced by the torque) and is positive in the direction of advance of a right-hand screw. Thus, the torque, in Fig. 1.21, being clockwise, is indicated by a vector pointing into the page away from the reader which is denoted by a cross in a circle; the cross representing the back of an arrow with its tip pointing into the page.

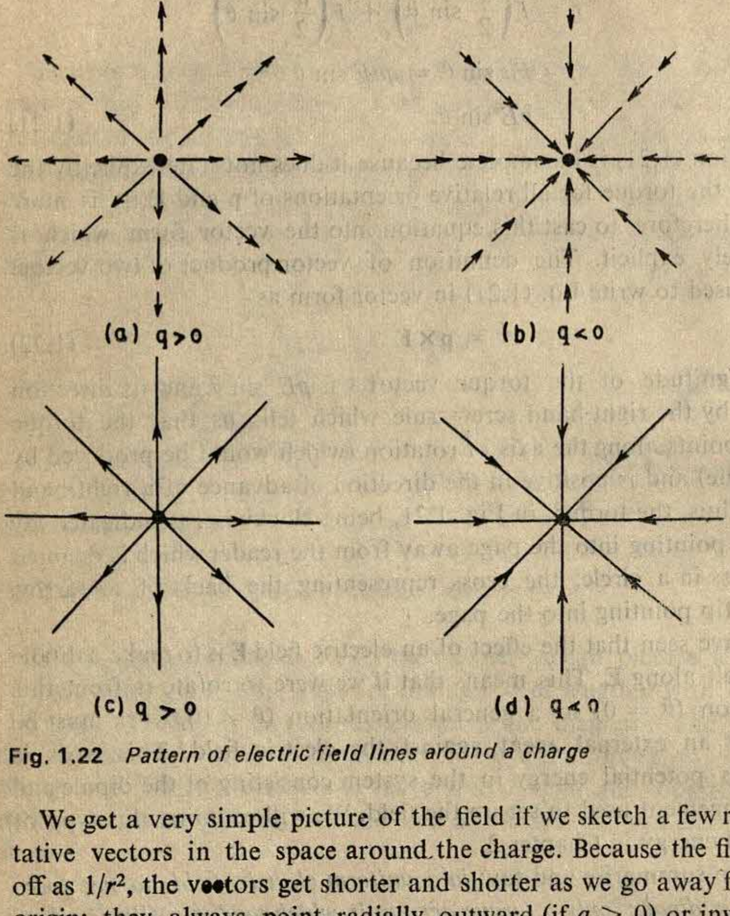
We have seen that the effect of an electric field **E** is to make a dipole align itself along **E**. This means that if we were to rotate it from this orientation ($\theta = 0$) to a general orientation ($\theta \neq 0$), work must be done by an external agent against the electric field. This work is stored as potential energy in the system consisting of the dipole and the arrangement used to set up the field. We will examine this type of potential energy in Chapter 2.

We can summarise and say: *in a uniform electric field no net force acts on a dipole; only a torque acts on it which tends to align it along the field.* To learn what happens to a dipole when it is placed in a nonuniform field, refer to the supplement on '*Dipole in a nonuniform field*' at the end of this chapter.

1.13 ELECTRIC LINES OF FORCE

The *lines of force of a field* (or *field lines*) is a very useful concept in physics. Lines of force are a pictorial representation of a field. Let us begin with the simplest example of an electric field due to a single point charge q placed, say, at the origin of the coordinate system in space. The electric field at a general point **r** away from O is given by

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \cdot \frac{q\hat{\mathbf{r}}}{r^2}$$



(a) $q > 0$ (b) $q < 0$

(c) $q > 0$ (d) $q < 0$

Fig. 1.22 *Pattern of electric field lines around a charge*

We get a very simple picture of the field if we sketch a few representative vectors in the space around the charge. Because the field falls off as $1/r^2$, the vectors get shorter and shorter as we go away from the origin; they always point radially outward (if $q > 0$) or inwards (if $q < 0$) as shown in Figs 1.22(a) and (b) respectively. The length of a vector at a point represents the magnitude of the force which a 'test' charge would experience if placed at that point. The direction in which the arrow points is the direction of the force (which is the same as that of the field) at that point. A much nicer way to represent this field is to connect up the arrows to form continuous lines, as shown in Figs 1.22(c) and (d). You might think that, by doing this, we have lost information about the strength of the field which was contained in the length of the vectors. But actually we have not. The magnitude of the field is indicated by the *density of the field lines*, the field is strong near the origin where the field lines are close together, and weak farther out, where they are far apart. Obviously, this field shows *radial* or *spherical symmetry*, i.e. it 'looks' the same in all directions, as seen from q.

Note: The field-line diagrams shown in Figs 1.22(c) and (d) are apparently misleading, when they are drawn on a two-dimensional surface (such as the plane of a paper). The density of lines passing through a circle of radius r is the total number (n) of lines divided by the circumference ($2\pi r$) is ($n/2\pi r$) which goes like ($1/r$) and not ($1/r^2$). But we imagine the model in three dimensions (a pincushion with needles sticking out in all directions) then the density of lines is the total number divided by the area of the sphere ($n/4\pi r^2$) which *does* go like ($1/r^2$).

Such diagrams are also convenient for representing more complicated fields such as those due two or more point charges, for which the field lines will, in general, be curved because the fields will not be radially symmetric unless the observation point is far away from the charges [Eqs (1.17) and (1.18) of a dipole field].

To draw a line of force in a region where an electric field exists, we proceed as follows. Take any point P in space where an electric field, say **E**, exists. Move from P by a small distance to a nearby point P_1 in the direction of **E**. Let the electric field at P_1 be $\mathbf{E}_1$. Then move from P_1 to a nearby point P_2 in the direction of $\mathbf{E}_1$ and so on from P_2 to P_3 to P_4, . . . in the direction of $\mathbf{E}_2$, $\mathbf{E}_3$, . . . (see Fig. 1.23). By bringing these points closer and closer, we get a continuous curve starting out at P in the direction of **E**.

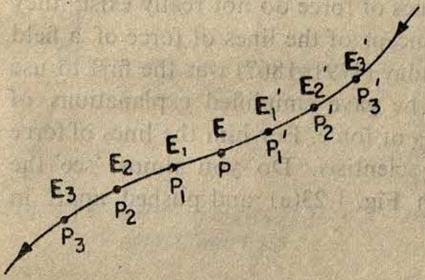


Fig. 1.23 *Constructing a line of force*

We can also go in the opposite direction from P to points P_1', P_2', P_3', . . . etc. so the curve extends on both sides of the starting point P. The point P is arbitrarily chosen; we could take any point on the curve as a starting point. The basic property of this curve is that the tangent to the curve at any point is parallel to the electric field at that point. Such a curve in space is called an *electric line of force* or *electric field line*.

In a similar way, any number of lines of force can be drawn in the region of the electric field. The number of lines should be enough to give an accurate description of the field. Figures 1.23(a) and (b) show the lines of force of the electric field due to (a) two equal unlike charges (i.e. a dipole) and (b) two equal like charges.

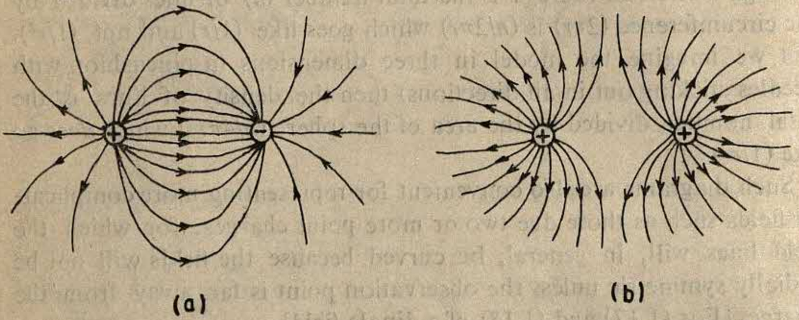


Fig. 1.24 *Lines of force for (a) two equal and opposite charges and (b) two equal positive charges*

The lines-of-force pattern helps us visualize the nature of the electric field (its sense and strength) in a region of space. They originate on a positive charge and terminate on a negative charge. It is easy to see that no two lines of force can ever cross because, if they did, the electric field would have two different directions at the point where they cross—which is not possible.

It may be mentioned that the lines of force do not really exist; they are hypothetical lines. Yet the concept of the lines of force of a field is a very useful one. Michael Faraday (1791–1867) was the first to use the idea of the lines of force and he gave simplified explanations of many phenomena in terms of lines of force. For him the lines of force were more real than for most scientists. Do you almost 'see' the charges being pulled together in Fig. 1.23(a) and pushed apart in Fig. 1.23(b) by the lines of force?

Properties of Lines of Force

The relationship between the lines of force and the electric field is obtained by the following properties of the lines of force:

1. *The tangent to a line of force at any point gives the direction of the field at that point.*

2. *The density of lines in a given region is proportional to the magnitude (strength) of the field in that region.* The density of lines means

the number of lines per unit area cutting a surface perpendicular to the direction of the lines at any given point.

3. *All lines originate from a positive charge and terminate on a negative charge.*

4. *The number of lines originating from or ending on a charge is proportional to the magnitude of the charge.* Thus if, say, 10 lines are shown originating from a charge q, then 20 lines should originate from a charge $2q$.

5. *No two lines of force ever cross each other.*

SUPPLEMENT

Electric Dipole in a Nonuniform Electric Field

A non-uniform electric field $\mathbf{E}(\mathbf{r})$ is the one that changes from point to point, either in magnitude or in direction or both. The lines of force of a nonuniform field are not parallel to each other. The field due to a single point charge or a dipole are nonuniform. When a dipole is placed in a nonuniform field; the expression for the torque on the dipole remains unchanged; it just depends on the location $\mathbf{r}$ of the dipole, the position $\mathbf{r}$ being the centre of the dipole.

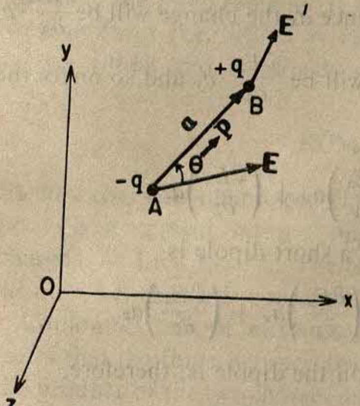


Fig. 1.25 *Force on a dipole in a nonuniform field*

$$\boldsymbol{\tau} = \mathbf{p} \times \mathbf{E}(\mathbf{r})$$

But now the net force $\mathbf{F}$ is *not* zero since the positive and negative charges are at different positions. Because the electric field is different at these positions, the two charges do not experience equal and opposite forces and hence a net force $\mathbf{F}$ acts on the dipole in a nonuniform field.

We shall first derive the expression for the x-component of the force $\mathbf{F}$. Let the negative charge $-q$ of the dipole be located at a point A in space whose position vector is $\mathbf{r}(x, y, z)$ while that of the positive charge at B is $\mathbf{r}'(x', y', z')$. The vector separation between the charges is $\mathbf{a}$ whose x, y and z components are a_x, a_y and a_z. Let E_x be the x-component of the electric field $\mathbf{E}$ at A. Let $\frac{\partial E_x}{\partial x}$, $\frac{\partial E_x}{\partial y}$ and $\frac{\partial E_x}{\partial z}$ be the rates at which E_x changes with x, y and z respectively. For a short dipole (where a is very small) these rates may be treated as nearly constant as we go from one end A of the dipole to the other end B.

Notice that we use partial derivatives $\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}$ rather than total derivatives $\frac{d}{dx}, \frac{d}{dy}, \frac{d}{dz}$ in accordance with the convention that if a quantity depends on more than one variable, we use partial derivatives. Here E_x depends upon three variable x, y and z, i.e. $E_x(x, y, z)$. The simple rule is: let y and z be fixed and treat E_x, for the moment, as a function of x alone, calculate $\partial E_x/\partial x$ and so on.

Now let us find the x-component of the electric field (E'_x) at the other end B of a very short dipole.

$$E'_x = E_x + \delta E_x$$

where δE_x is a small change in E_x as we go from A to B. Remember ∂E_x is also a function of three variables x, y and z. Since $\partial E_x/\partial x$ is the rate (treated constant as a first approximation over a short dipole) at which E_x changes with x, in a distance a_x the change will be $\frac{\partial E_x}{\partial x} \cdot a_x$

Similarly in a distance a_y the change will be $\frac{\partial E_x}{\partial y} \cdot a_y$ and so on for the z-direction. Hence

$$\delta Ex = \left(\frac{\partial E_x}{\partial x}\right)a_x + \left(\frac{\partial E_x}{\partial y}\right)a_y + \left(\frac{\partial E_x}{\partial z}\right)az$$

Thus the electric field at the end B of a short dipole is

$$E'_x = E_x + \left(\frac{\partial E_x}{\partial x}\right)a_x + \left(\frac{\partial E_x}{\partial y}\right)a_y + \left(\frac{\partial E_x}{\partial z}\right)a_z$$

The x-component of the net force **F** on the dipole is, therefore,

$$\begin{aligned} F_x &= -qE_x + qE'_x \\ &= q\left[a_x\left(\frac{\partial E_x}{\partial x}\right) + a_y\left(\frac{\partial E_x}{\partial y}\right) + a_z\left(\frac{\partial E_x}{\partial z}\right)\right] \end{aligned}$$

Since the dipole moment $\mathbf{p} = q\mathbf{a}$, we have

$$F_x = p_x\left(\frac{\partial E_x}{\partial x}\right) + p_y\left(\frac{\partial E_x}{\partial y}\right) + p_z\left(\frac{\partial E_x}{\partial z}\right)$$

or

$$F_x = \left(p_x\frac{\partial}{\partial x} + p_y\frac{\partial}{\partial y} + p_z\frac{\partial}{\partial z}\right)E_x \tag{1.23}$$

where p_x, p_y and p_z are the components of **p** along the three axes.

We can now easily obtain the expressions for the y and z components of $\mathbf{F}$. To obtain F_y we simply replace E_x in Eq. (1.23) by E_y and to get F_z we replace E_x by E_z. Thus

$$F_y = \left(p_x \frac{\partial}{\partial x} + p_y \frac{\partial}{\partial y} + p_z \frac{\partial}{\partial z}\right) E_y \tag{1.24}$$

and

$$F_z = \left(p_x \frac{\partial}{\partial x} + p_y \frac{\partial}{\partial y} + p_z \frac{\partial}{\partial z}\right) E_z \tag{1.25}$$

Equations (1.23), (1.24) and (1.25) can be written in a compact vector form as

$$\mathbf{F} = p\left(\hat{p}_x \frac{\partial}{\partial x} + \hat{p}_y \frac{\partial}{\partial y} + \hat{p}_z \frac{\partial}{\partial z}\right) \mathbf{E} \tag{1.26}$$

where $\hat{p}_x$, $\hat{p}_y$ and $\hat{p}_z$ are the three components of the unit vector $\hat{\mathbf{p}} = \mathbf{p}/p$ along the three axes.

Thus *a dipole in a nonuniform field experiences a nonzero force and a nonzero torque*. In the special case when the dipole is parallel to the field ($\theta = 0$) (see Fig. 1.24) the torque becomes zero (since $\tau = pE \sin \theta$) and only the force remains. It is clear from Eq. (1.26) that if the field were uniform (i.e. $\partial \mathbf{E}/\partial x = \partial \mathbf{E}/\partial y = \partial \mathbf{E}/\partial z = 0$), the force on the dipole will be zero, as already seen in Sec. 1.12.

SUMMARY

Electric charge is quantized. The magnitude of charge on a charged body is a certain integral multiple of the fundamental minimum charge $e = 1.6 \times 10^{-19}$ C. The total charge of the universe is fixed for all time, i.e. charge is conserved.

Coulomb's law of electrostatic force between two point charges states that the force between two charges is directly proportional to the product of the two charges and inversely proportional to the square of the distance between them.

The principle of superposition states that the total force on a given charge is the vector sum of the individual forces exerted on it by all other charges; the force between two charges being completely unaffected by the presence of other charges.

A pair of equal and opposite point charges ($\pm q$) separated by a vector distance $\mathbf{a}$ is called a dipole. The quantity $\mathbf{p} = q\mathbf{a}$ is the dipole moment. When a dipole is placed in a uniform electric field, it experiences no net force but a nonzero torque. The torque exerted on the dipole by the field tends to align it parallel to the field. When a dipole is placed in a nonuniform electric field, it experiences a nonzero force and a nonzero torque. In the special case when the dipole moment is parallel to the field, the dipole experiences a zero torque but a nonzero force. The torque is given by $\boldsymbol{\tau} = \mathbf{p} \times \mathbf{E}$; its SI unit is N m; the SI units of **p** and **E** are Cm and NC^{-1} respectively.

The line of force of a field is a curve, the tangent to which at any point gives the direction of the field at that point. The density of lines of force in a given region is a measure of the magnitude of the field in that region.

EXERCISES

A. Short Answer Questions

1. 'Electric charge is quantized'. What does this statement mean?
2. What do you understand by 'conservation of electric charge'?
3. State Coulomb's law of electrostatic force.
4. State, with unit, the value of permittivity of free space.
5. Define a coulomb of charge.
6. Distinguish between a conductor and an insulator. Give one example of each.
7. State the principle of superposition in relation to forces acting on a point charge.
8. 'The force between two charges is completely unaffected by the presence of other charges'. Comment.
9. What do you understand by electric field? What is its SI unit?
10. What is an electric dipole? Define dipole moment.
11. 'A dipole in a uniform field experiences no net force'. Is this statement true or false? Give the reason for your answer.
12. 'The torque exerted on a dipole in a nonuniform field is independent of the orientation of the dipole with respect to the field'. Is this statement true or false? Give the reason for your answer.
13. Will a dipole have a translational motion when it is placed in a nonuniform field? Give the reason for your answer.
14. State two properties of the lines of force of a field.
15. No two lines of force ever cross each other. Why?

B. Long Answer Questions

1. How will you explain electrification by friction in terms of the atomic theory.
2. How can a charged body (a plastic comb rubbed with dry hair) attract an uncharged body (a tiny piece of paper)? Explain, describing the phenomenon involved.
3. Describe (with explanation) the two methods of charging conductors.
4. State the principle of superposition and use it to obtain the expression for the total force exerted on a point charge due to an assembly of N discrete point charges.

5. Name the three kinds of continuous charge distributions and write down the expression for the force exerted on a test charge by each distribution.
6. What do you understand by the term 'electric field'? How is it defined and measured? What is the usefulness and physical significance of this concept?
7. Obtain the expression for the electric field of a dipole at a point: (a) on its axis and (b) on the perpendicular bisector of the line joining the two charges. Discuss the limiting case of a point far away from the dipole. How does the electric field in this case compare with that of a single point charge?
8. Deduce the expression for the electric field at a point on the axis of a conducting charged rod of negligible thickness. Hence show that, for points far away, the rod behaves as though it were a point charge.
9. A circular loop of wire of radius r and negligible thickness carries a charge q which is distributed uniformly over the loop. Find the electric field at a point situated at a distance x from the centre along the line perpendicular to the plane containing the loop and passing through its centre.
10. An electric dipole is placed at rest in a uniform external electric field. Discuss its motion.
11. Find the net force acting on a short dipole placed in a nonuniform electric field.
12. Define the line of a force of a field. How is this concept useful? State the properties of lines of force.
13. Show, in a diagram, the pattern of the lines of force in the space around: (a) a positive point charge, (b) a pair of equal positive charges and (c) an electric dipole.
14. If a point charge q of mass m is released from rest in a nonuniform electric field, will it follow a line of force? Give the reason for your answer.

(C) Multiple Choice Questions

Choose the correct answer or answers from the following:

1. The SI unit of electrical permittivity is
   (a) $C^2 N^{-1} m^{-2}$ (b) $N m^2 C^{-2}$
   (c) $F m^{-1}$ (d) $N C^{-1}$
2, What is the SI unit of electric field?
   (a) $A m^{-1}$ (b) Vm^{-1}
   (c) $C m$ (d) NC^{-1}
3. Two point-charges, 3.0×10^{-8} C and 8.0×10^{-8} C, repel each other with a force of 6×10^{-3} N. What is the force between them if each is given an additional charge of -6.0×10^{-8} C?
   (a) 1.5×10^{-3} N (attractive)
   (b) 1.5×10^{-3} (repulsive)
   (c) 2.4×10^{-2} N (attractive)
   (d) 2.4×10^{-2} N (repulsive)
4. Three point-charges $4q$, Q and q, are placed in a straight line of length l at points distant 0, $l/2$ and l respectively. What must be the value of Q so that the net force on charge q is zero?
   (a) $-q$ (b) $-2q$
   (c) $-\frac{1}{2}q$ (d) $4q$

5. Which of the patterns shown in Fig. 1.26 *cannot* possibly represent electrostatic field lines?

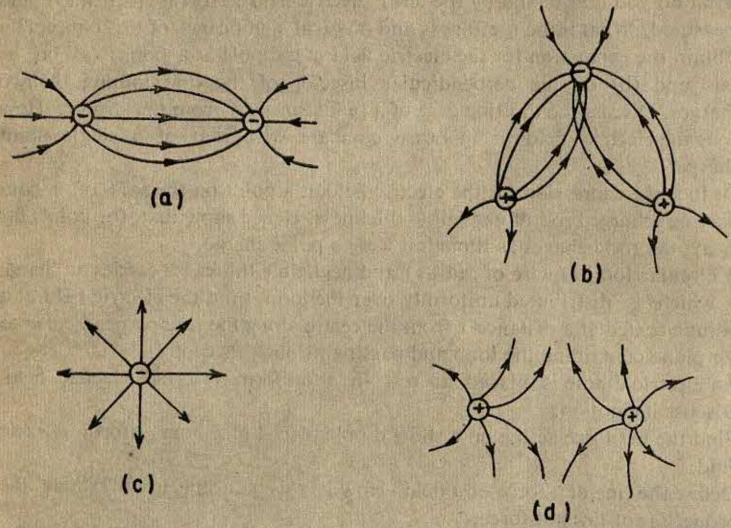


Fig. 1.26

6. When an electric dipole is placed in a uniform electric field, it experiences
   (a) a force as well as a torque
   (b) a torque but no force
   (c) neither a force nor a torque
   (d) a force but no torque.
7. When an electric dipole is placed in a nonuniform field, it experiences
   (a) a force as well as a torque
   (b) neither a force nor a torque
   (c) a torque but no force
   (d) a force but no torque.
8. When an electric dipole is placed in a uniform electric field with its dipole moment parallel to the field, it will experience
   (a) a torque but no force
   (b) a force but no torque
   (c) a torque as well as a force
   (d) neither a force nor a torque.

D. Numerical Problems

(Take $\epsilon_0 = 8.85\times10^{-12}\ \mathrm{C^2N^{-1}m^{-2}}$, $e = 1.6\times10^{-19}$ C and $g = 9.8\ \mathrm{ms^{-2}}$ wherever necessary.)

1. An alpha particle is a helium nucleus having two protons and two neutrons. Calculate the electrostatic force between two alpha particles separated by a distance of 10^{-13} m.

2. Three charges $q_1 = -1.0\ \mu C$, $q_2 = +3.0\ \mu C$ and $q_3 = -2.0\ \mu C$ are placed as shown in Fig. 1.27. What is the force acting on q_1?

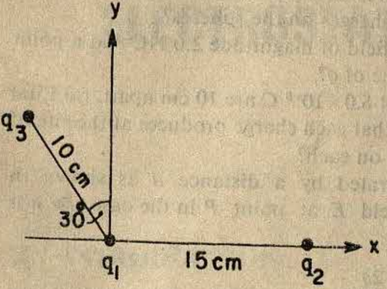


Fig. 1.27

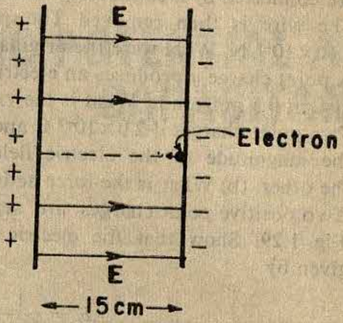


Fig. 1.28

3. Two small identical balls are suspended on threads so that their surfaces are in contact. A charge of 4×10^{-7} C is given to each ball which makes them repel each other to an angle of 60° between the threads. Find the mass of each ball if the distance from the point of suspension to the centre of a ball is 20 cm.

4. Two charged conducting plates are 15 cm apart in vacuum (Fig. 1.28). The electric field between the plates has a magnitude 3000 NC^{-1}. An electron is released from rest at a point close to the negative plate. How long will it take to reach the other plate? What is its speed just before it hits the plate?

5. Two small identical balls, each of mass m, are hung from a common point by silk threads of length l and carry similar charges. At equilibrium, the angle between the threads is 2θ and the separation between the balls is x. Assuming that θ is so small that tan θ can be replaced by sin θ show that

$$x = \left(\frac{q^2 l}{2\pi\epsilon_0 mg}\right)^{1/3}$$

If $l = 100$ cm, $m = 5.0$ g, what value of q will produces a separation

$$x = 2.5 \text{ cm?}$$

6. The electrostatic force of repulsion between two positively charged ions, carrying equal charge, is 3.7×10^{-9} N when they are separated by a distance of 5.0×10^{-10} m. How many electrons are missing from each ion?

7. Two small positively charged spheres have a combined charge of 5.00×10^{-5} C. When they are 2.00 m apart, the force of repulsion between them is 1.00 N. Find the charge on each sphere.

8. Find the magnitude and the direction of the electric field at the centre of a square 2 m on each side, if three charges of $+8\ \mu C$ each are placed at three of the corners of the square.

9. Two equal point charges $Q = +\sqrt{2}\ \mu C$ are placed at each of the two opposite corners of a square. A point charge q is placed at each of the other two corners. What must be the value of q such that the resultant force on charge Q is zero?

10. Two oppositely charged small copper spheres of the same size attract each other with a force of 1.08×10^{-1} N when separated by 50 cm. The spheres are connected by a conducting wire without their locations being changed. The wire is then removed. The spheres now repel each other by a force of 3.60×10^{-2} N. What were the original charges on the spheres?
11. A point charge q produces an electric field of magnitude 2.0 NC^{-1} at a point distant 0.5 m from it. What is the value of q?
12. Two point charges $+2.0 \times 10^{-8}$ C and $+8.0 \times 10^{-9}$ C are 10 cm apart. (a) Find the magnitude of the electric field that each charge produces at the site of the other. (b) What is the force acting on each?
13. Two positive point charges are separated by a distance a as shown in Fig. 1.29. Show that the electric field E at point P in the case $r \gg a$ is given by

$$E = \frac{1}{4\pi\epsilon_0} \cdot \frac{2q}{r^2}$$

What is the direction of **E**. Compare this result with that of a dipole field for large value of r.

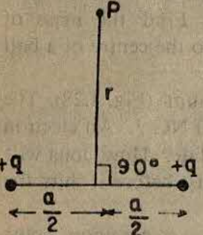


Fig. 1.29

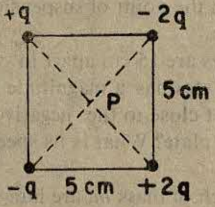


Fig. 1.30

14. Four point charges are placed at the corners of a square of side 5.0 cm as shown in Fig. 1.30. If $q = 1.0 \times 10^{-8}$ C, find the magnitude and the direction of the electric field at the intersection P of the diagonals.
15. What is the magnitude and direction of an electric field that will balance an alpha particle?
16. A particle having a charge of -2.0×10^{-9} C experiences a downward electric force of 3.0×10^{-6} N in a uniform electric field. What is the magnitude and direction of the force which a proton will experience in that field?
17. Two identically charged spheres are suspended by strings of equal lengths. The strings make an angle of 30° with each other. When suspended in a liquid of density 800 kg m^{-3}, the angle remains the same. What is the dielectric constant of the liquid? The density of the material of the spheres is 1600 kg m^{-3}.
18. A charge of $+3\ \mu\text{C}$ is placed at $x = 0$ and a charge of $-5\ \mu\text{C}$ at $x = 40$ cm. Where must a third charge q be placed on the x-axis such that it experiences no force?
19. Equal charges of $+4\ \mu\text{C}$ are placed at the three corners of an equilateral triangle of side 2 m. Calculate the magnitude and direction of the force on one of the charges.

2

ELECTROSTATIC POTENTIAL AND CAPACITANCE

2.1 INTRODUCTION

In the previous chapter we learnt about forces between charges and introduced the very important concept of the electric field due to an isolated point charge or a distribution of charges. We have seen how the electric field around a charge can be described in terms of a *vector* quantity **E**. In this chapter we will first learn how the electric field around a charge can also be described by a *scalar* quantity ϕ called the *electrostatic* or *electric potential.* The close relationship between potential and electric field is demonstrated, and some examples are developed which show the usefulness of the potential as a means for learning about the electric field.

The concept of electric potential can be best understood in terms of a very important property of an electrostatic field, namely, that it is *conservative*. You have already learnt in Class XI that the gravitational field is conservative, i.e. the work done in climbing a hill does not depend upon the route we take; it depends only on the starting and end points of the route. This is perhaps expected since both Coulomb's law and the law of gravitation obey the inverse square dependence on distance.

2.2 AN IMPORTANT PROPERTY OF ELECTROSTATIC FIELDS

Let us suppose that there is a system of electric charges (an assembly of discrete point charges or continuous charge distributions) which produces an electric field **E** in the space surrounding the system of charges. Consider any two points A and B in space and connect them by a continuous curve L running from A to B. Let us construct the line integral of **E** from A to B along L (refer to 'line integral of vector fields' in Appendix I at the end of the book). We now ask the question: Will the value of the line integral depend upon the path (or route) that we take to go from A to B (of course keeping A and B fixed)? We show in the following that it will not; the line integral will simply depend upon the starting and end points A and B and not on L. This is the basic property of electrostatic fields which we state explicitly first and then proceed to prove it:

For electrostatic fields, the line integral

$$\int_{\substack{A \\ \text{along L}}}^{B} \mathbf{E}\cdot d\mathbf{l} \text{ is the same for all paths L, depending only on the end points A and B.} \quad (2.1)$$

We will establish this property first for the electric field produced by a stationary point charge. For convenience, we assume that the charge is located at the origin of the coordinate system in space. We choose any two points A and B and connect them by the curve L. Let $\mathbf{r}_A$ and $\mathbf{r}_B$ be the position vectors of A and B respectively (Fig. 2.1).

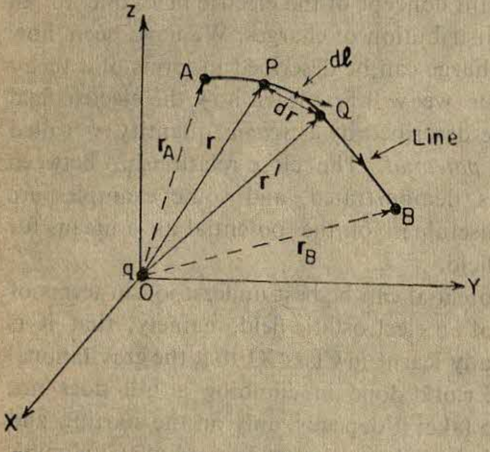


Fig. 2.1 *Line integral of the electric field of a point charge*

We will now evaluate the line integral (2.1). We know that electric field **E** due to a point charge q at a point **r** is given by

$$\mathbf{E} = \frac{1}{4\pi\epsilon_0}\cdot\frac{q}{r^2}\cdot\hat{\mathbf{r}} = \frac{q}{4\pi\epsilon_0}\cdot\frac{\mathbf{r}}{r^3}$$

Thus the line integral to be evaluated is

$$\int_{\substack{A \\ \text{along L}}}^{B} \mathbf{E}\cdot d\mathbf{l} = \frac{q}{4\pi\epsilon_0}\int_{\substack{A \\ \text{along L}}}^{B} \frac{\mathbf{r}\cdot d\mathbf{l}}{r^3} \quad (2.2)$$

In order to evaluate the integral in Eq. (2.2) we divide the curve L into a very large (infinite) number of very small (infinitesimal) segments such as PQ (which has been exaggerated in the diagram for convenience). Let **r** and **r**′ be the position vectors of P and Q respectively. It is clear from Fig. 2.1 that the vector d**r** is given by

$$\mathrm{d}\mathbf{r} = \mathbf{r}' - \mathbf{r}$$

or
$$\mathbf{r}' = \mathbf{r} + \mathrm{d}\mathbf{r} \tag{2.3}$$

Now, as point Q moves closer and closer to point P, the vector d**r** becomes more and more equal to the vector d**l** representing the line element and in the limit Q → P,

$$\mathrm{d}\mathbf{r} = \mathrm{d}\mathbf{l}$$

so that
$$\mathbf{r}' = \mathbf{r} + \mathrm{d}\mathbf{l} \tag{2.4}$$

It must be remembered that the line element d**l** is so small the square of its magnitude (which will be even smaller) can be neglected. To this approximation, we have, from Eq. (2.4).

$$\begin{aligned} r'^2 &= \mathbf{r}'\cdot\mathbf{r}' = (\mathbf{r} + \mathrm{d}\mathbf{l})\cdot(\mathbf{r} + \mathrm{d}\mathbf{l}) \\ &= \mathbf{r}\cdot\mathbf{r} + 2\mathbf{r}\cdot\mathrm{d}\mathbf{l} + \mathrm{d}\mathbf{l}\cdot\mathrm{d}\mathbf{l} \\ &= r^2 + 2\mathbf{r}\cdot\mathrm{d}\mathbf{l} + (\mathrm{d}l)^2 \\ &\simeq r^2 + 2\mathbf{r}\cdot\mathrm{d}\mathbf{l} \end{aligned}$$

or
$$\begin{aligned} \mathbf{r}\cdot\mathrm{d}\mathbf{l} &= \frac{1}{2}(r'^2 - r^2) \\ &= \frac{1}{2}(r' + r)(r' - r) \end{aligned}$$

Now $r' = r + \mathrm{d}r$ where $\mathrm{d}r$ is small and its square can be neglected to our approximation. Therefore

$$\begin{aligned} \mathbf{r}\cdot\mathrm{d}\mathbf{l} &= \frac{1}{2}(r + \mathrm{d}r + r)(r + \mathrm{d}r - r) \\ &= \left(r + \frac{1}{2}\mathrm{d}r\right)\mathrm{d}r \end{aligned}$$

or
$$\mathbf{r}\cdot\mathrm{d}\mathbf{l} \simeq r\,\mathrm{d}r \tag{2.5}$$

Using Eq. (2.5) in (2.2) we get

$$\begin{aligned} \int_{\substack{A \\ \text{along L}}}^{B} \mathbf{E}\cdot\mathrm{d}\mathbf{l} &= \frac{q}{4\pi\epsilon_0}\int_{\substack{A \\ \text{along L}}}^{B} \frac{\mathrm{d}r}{r^2} \\ &= -\frac{q}{4\pi\epsilon_0}\cdot\frac{1}{r}\Bigg|_{r_A}^{r_B} = \frac{q}{4\pi\epsilon_0}\left(\frac{1}{r_A} - \frac{1}{r_B}\right) \end{aligned} \tag{2.6}$$

where r_A and r_B are the distances of points A and B from the point charge q (at origin). This result shows that the *line integral as we move from A to B is independent of the particular path L taken and depends only the positions of A and B*, the end points of the path. In fact, all that matters is how far A and B are from the charge located at the origin. The same result (2.6) holds if the charge is located at any general point rather than at the origin.

We have seen that the line integral of the electric field due to a single stationary charge is independent of the path taken. We can easily extend this result to an arbitrary electrostatic field due to any system of charges. It follows from the principle of superposition that the electric field at a point, due to any system of charges, is the vector sum of many Coulomb like contributions, one from each charge, each of which obeys the result (2.6). Since each term is independent of the path taken, the sum (i.e. superposition) of all the terms is also independent of the path. This means that the general result (2.1) is true for any arbitrary electrostatic field. Thus the basic property of any general electrostatic field is that *the line integral of the electrostatic field is path-independent, depending only on the end-points of the path.*

Note: The result (2.1) does not hold if the electric field is not electrostatic, e.g. fields due to moving charges. Thus if the electric field were varying with time or were a function of the velocity of charges, the line integral of such a field would not be independent of the path. We will come across such more general electric fields (other than electrostatic ones) in Chapter 7.

There is an equivalent way in which we can express the content of our basic result (2.1). *The line integral of an electrostatic field around a closed path is zero*, i.e.

$$\oint \mathbf{E}\cdot d\mathbf{l} = 0 \tag{2.7}$$

where we have used a special symbol (a circle on the integral sign) which represents integration around a closed path. This can be understood as follows:

Consider a closed path (a loop) in space as shown in Fig. 2.2. The loop starts at a point A and finally ends at A itself. Let us divide the total path into two parts L_1 (from A to a point, say B) and L_2 (from B back to A). Then from Eq. (2.6) we have

$$\underset{\text{along } L_1}{\int_A^B} \mathbf{E}\cdot d\mathbf{l} = \frac{q}{4\pi\epsilon_0}\left(\frac{1}{r_A} - \frac{1}{r_B}\right)$$

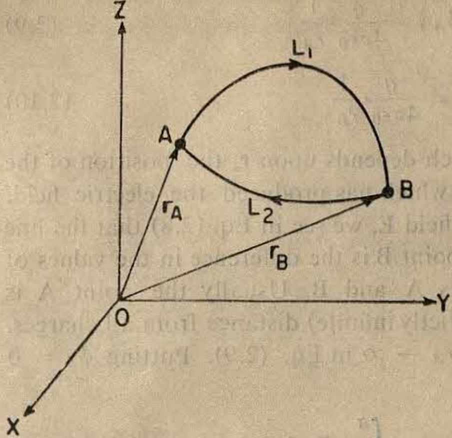


Fig. 2.2 *Line integral of an electric field around a closed path is zero*

and
$$\int_{\substack{B \\ \text{along } L_2}}^{A} \mathbf{E}\cdot d\mathbf{l} = \frac{q}{4\pi\epsilon_0}\left(\frac{1}{r_B} - \frac{1}{r_A}\right)$$
$$= -\frac{q}{4\pi\epsilon_0}\left(\frac{1}{r_A} - \frac{1}{r_B}\right)$$
$$= -\int_{\substack{A \\ \text{along } L_1}}^{B} \mathbf{E}\cdot d\mathbf{l}$$

Thus
$$\oint_{\text{along loop}} \mathbf{E}\cdot d\mathbf{l} = \int_{\substack{A \\ \text{along } L_1}}^{B} \mathbf{E}\cdot d\mathbf{l} + \int_{\substack{B \\ \text{along } L_2}}^{A} \mathbf{E}\cdot d\mathbf{l}$$
$$= \int_{A}^{B} \mathbf{E}\cdot d\mathbf{l} - \int_{A}^{B} \mathbf{E}\cdot d\mathbf{l} = 0$$

which is our result (2.7). This result holds for a loop starting from any point A, going out anywhere in space and finally returning to the same point A.

2.3 ELECTROSTATIC POTENTIAL AND POTENTIAL DIFFERENCE

The fact that the line integral of an electrostatic field is independent of the path leads us to a physically important and useful concept—the electrostatic potential. This fact is expressed by Eq. (2.6) which may be rewritten as

$$\int_{A}^{B} \mathbf{E}\cdot d\mathbf{l} = (\phi_A - \phi_B) \tag{2.8}$$

where $$\phi_A = \frac{q}{4\pi\epsilon_0} \cdot \frac{1}{r_A} \tag{2.9}$$

and $$\phi_B = \frac{q}{4\pi\epsilon_0} \cdot \frac{1}{r_B} \tag{2.10}$$

Here ϕ is a scalar function which depends upon **r**, the position of the point and q, the source charge which has produced the electric field. Thus, given some electrostatic field **E**, we see in Eq. (2.8) that the line integral of **E** from point A to point B is the difference in the values of some scalar function ϕ at points A and B. Usually the point A is chosen to be at a very large (strictly infinite) distance from all charges. With this choice $\phi_A = 0$ since $r_A \to \infty$ in Eq. (2.9). Putting $\phi_A = 0$ in Eq. (2.8) we have

$$\phi_B = -\int_{\infty}^{B} \mathbf{E} \cdot d\mathbf{l} \tag{2.11}$$

The scalar function ϕ defined in this manner is the *electrostatic potential* (or *electric potential*) corresponding to the field **E**, ϕ_B represents potential at point B. Remember the value of ϕ depends on the point chosen (and of course the source charge) in the space.

The potential difference between two points A and B is often written as ϕ_{AB} or $(\phi_B - \phi_A)$. From Eq. (2.8) we notice that

$$\phi_B - \phi_A = -\int_{A}^{B} \mathbf{E} \cdot d\mathbf{l} \tag{2.12}$$

i.e. *the line integral of an electrostatic field between any two points equals the negative of the potential difference between them.*

It is evident from Eq. (2.12) that the dimensions of ϕ are dimensions of $E \times$ dimension of dl, i.e. $NC^{-1} \times m = Nm\,C^{-1} = JC^{-1}$. In the SI system JC^{-1} is called *volt* in honour of the Italian scientist Alessandro Volta. Thus

$$1 \text{ volt} = 1\ Nm\,C^{-1} = 1 JC^{-1} \tag{2.13}$$

So, in the SI system, the unit of electric potential (and hence of potential difference) is the *volt* (symbol V).

Physical Meaning of Potential and Potential Difference

We can give a physical meaning to potential and potential difference in terms of the work required to be done on a charge in moving it from one place to another in an electrostatic field. Let us consider that an arbitrary assembly of static 'source' charges produces an electric field in a region of space. We will calculate the work necessary to be done by an external agent to move a 'test' charge q_0 (small and positive) from a position A to a position B against the force of an electric field E along a path as shown in Fig. 2.3.

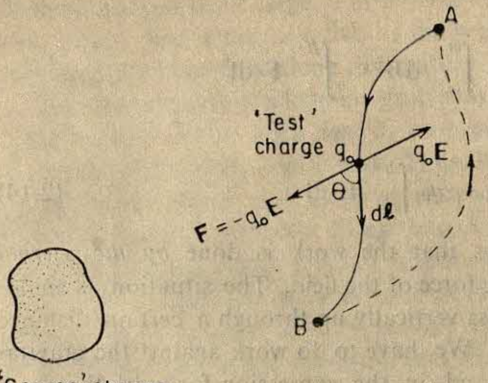


Fig. 2.3 *A test charge q_0 moved from A to B in a nonuniform electrostatic field* E *by an external agent that exerts a force* F

Let **E** be the electric field at the site of the test charge q_0. The electric field exerts a force q_0 **E** on the test charge as shown. To keep the test charge in equilibrium (i.e. to prevent it from accelerating) the external agent must exert a force **F** which must be enough to overcome the repulsive force q_0E exerted by the field on the test charge, for all positions of this charge, i.e.

$$\mathbf{F} = -q_0\mathbf{E}$$

Remember that **E** (and therefore **F**) are changing from place to place. The displacement of the charge from A to B is performed in such a way that the external force is directed opposite to the force of the field at all positions of the test charge. Under this condition all the work done by the external force **F** is in moving the charge from its original position A to its final position B and no part of it contributes to kinetic energy because the charge is always in equilibrium and hence does not accelerate.

To find out the total work W done by the external agent in moving the test charge from A to B, we divide the path into a large number of infinitesimal elements **dl** and find out the infinitesimal work dW done on the charge to move it through **dl**. Since the force **F** and displacement **dl** are inclined to each other at an angle θ (see Fig. 2.3), the work dW is given by the dot product of the two vectors **F** and **dl**, i.e.

$$\mathrm{d}W = \mathbf{F}\cdot\mathbf{dl}$$

The total work W is obtained by adding up (i.e. integrating) the work contributions dW for all the infinitesimal elements into which the path is divided. Thus

$$W_{AB} = \int_A^B dW = \int_A^B \mathbf{F}\cdot d\mathbf{l}$$

But $\mathbf{F} = -q_0\mathbf{E}$. Hence,

$$W_{AB} = -q_0 \int_A^B \mathbf{E}\cdot d\mathbf{l} \qquad (2.14)$$

The *negative* sign indicates that the work is done *by the external agency* against the repelling force of the field. The situation is analogous to that of lifting a mass vertically up through a certain distance above the earth's surface. We have to do work against the gravitational pull of the earth and in the expression for work there is a negative sign in front of the product of the gravitational force and the distance through which the mass is raised. (Refer to the chapter on gravitation that you have studied in Class XI.) In our case also [see Eq. (2.14)] we have a negative sign in front of the product of electrostatic force $q_0\mathbf{E}$ of the field and the displacement $d\mathbf{l}$ of the charge q_0.

In terms of our definition of potential difference [Eq. (2.8)], Eq. (2.14) reads

$$W_{AB} = q_0(\phi_B - \phi_A)$$

or

$$(\phi_B - \phi_A) = \frac{W}{q_0} \qquad (2.15)$$

i.e.

$$\text{Potential difference} = \frac{\text{work done}}{\text{test charge}}$$

We now have a physical interpretation of potential difference. *The potential difference between two points in an electrostatic field is the amount of work per unit charge done by the external agent to move the small charge from one point to the other against the electrostatic force of the field along any path.* Note the phrase '*along any path*'. We have already seen that the line integral in Eq. (2.14) is independent of the path. Hence the work done in moving a charge from one point to another in an electrostatic field is independent of the path. In Fig. 2.3 only one path is shown for convenience. In other words, the *electrostatic field is conservative.* You have learnt in Class XI that the gravitational field is also conservative. The amount of work we have to do to lift a mass from one point to another against the gravitational force of the earth does not depend upon what route we take; it depends only on the end points of the route.

We have seen that the line integral of an electric field has a simple interpretation in terms of work. Equation (2.14) tells us that the work done on a small test charge q_0 in moving it along any path against the electrostatic field is the negative of the product of the charge and the line integral of the field along that path.

From the property of a loop integral in an electrostatic field [Eq. (2.7)], it follows that *the net work done in carrying a charge over a closed loop in an electrostatic field is zero*. Thus, if we take a charge from A to B and then return to A along an alternate path (shown dashed in Fig. 2.3), the work done in going from A to B must be just the negative of the work from B to A; so that the net work around any closed path must be zero. The work is done *on the charge* to move from A to B, an equal amount of work is done *by the charge* to move from B to A. We will see in the following that the work done on a charge is stored as the *potential energy* of the charge. We will see in Chapter 7 that a general time-dependent electric field is not conservative.

Definition of Electric Potential

If the point A is taken to be at infinity it follows from Eq. (2.9) that $\phi_A = 0$, in which case, Eq. (2.15) becomes (dropping the subscript B)

$$\phi = \frac{W}{q_0} \tag{2.16}$$

Thus, *the electric potential at a point in an electrostatic field is the work per unit charge that is done to bring a small charge in from infinity to that point along any path*. This interpretation as work done per unit charge explains why the volt is the same as joule per coulomb, i.e.

$$1\text{V} = 1\text{JC}^{-1}$$

as stated earlier.

3.4 POTENTIAL DUE TO SOME CHARGE DISTRIBUTIONS

We will now illustrate how we can calculate electric potential due to some charge distributions.

1. Potential due to a Point Charge

We have already obtained the expression for the potential at a point P in the electrostatic field of a point charge q. It is

$$\phi_P = \frac{1}{4\pi\epsilon_0}\cdot\frac{q}{r} \tag{2.17}$$

where r is the distance of the point P from the charge. This potential is spherically symmetric around the point, i.e. it depends only on r for a given charge q. Since potential is a scalar function, the spherical symmetry means that the potential at a point does not depend upon the direction of that point with respect to the point charge; it only depends on the distance of the point from the charge.

Notice from Eq. (2.17) that the potential due to a positive charge ($q > 0$) is positive, it is negative in the neighbourhood of an isolated negative charge ($q < 0$). This is expected because if we bring a small positive charge q_0 from infinity to a distance r from q, we would have to overcome the force of repulsion that q exerts on q_0, and so we would indeed have to do a positive amount of work.

2. Potential due to a System of Charges

When we consider the electric field due to more charges than one, we see the advantages of the concept of potential over that of electric field. The reason is that, like electric field, potential also obeys the superposition principle. Thus, to find the electric potential at a point in the electric field due to two or more charges, we first calculate the potential due to each charge, assuming that all other charges are absent, and then simply add these individual contributions. Since, unlike electric field, electric potential is a scalar, the addition here is the *ordinary* sum, not a *vector* sum. Therefore, it is much easier to work with potential than with electric field.

(i) *Two Point Charges* The potential at any point due to two point charges q_1 and q_2 is, therefore, simply given by

$$\phi = \frac{1}{4\pi\epsilon_0}\left(\frac{q_1}{r_1} + \frac{q_2}{r_2}\right) \tag{2.18}$$

where r_1 and r_2 are the distances of the point in question from charges q_1 and q_2 respectively.

(ii) *Many Point Charges* The potential at any point due to a system of N discrete point charges is given by

$$\phi = \sum_{n=1}^{N} \phi_n = \frac{1}{4\pi\epsilon_0} \sum_{n=1}^{N} \frac{q_n}{r_n} \tag{2.19}$$

where $n = 1, 2, 3, \ldots, N$ and ϕ_n is the potential due to the nth charge q_n (as if the other charges were absent) are r_n is the distance of this charge from the point where the potential ϕ is to be found.

(iii) *Continuous Charge Distributions* If the charge distribution is continuous, the summation in Eq. (2.19) is replaced by integration (see Sec. 1.10).

$$\phi = \frac{1}{4\pi\epsilon_0}\int \frac{\mathrm{d}q}{r} \tag{2.20}$$

where dq is a differential element of the charge distribution and r its distance from the point at which ϕ is to be calculated.

If the charge is continuously distributed inside a volume V, then $\mathrm{d}q = \rho\,\mathrm{d}V$ where ρ is volume density of charge and the integral in Eq. (2.20) is the volume integral. Thus

$$\phi_v = \frac{1}{4\pi\epsilon_0}\int_V \frac{\rho\,\mathrm{d}V}{r} \tag{2.21}$$

If the charge is spread continuously over an area S, then $\mathrm{d}q = \sigma\,\mathrm{d}S$, where σ is the surface density of charge and the integral in Eq. (2.20) is the surface integral, i.e.

$$\phi_S = \frac{1}{4\pi\epsilon_0}\int_S \frac{\sigma\,\mathrm{d}S}{r} \tag{2.22}$$

If the charge is uniformly distributed along a line l, then $\mathrm{d}q = \lambda\,\mathrm{d}l$, where λ is the linear charge density and we have

$$\phi_l = \frac{1}{4\pi\epsilon_0}\int_l \frac{\lambda\,\mathrm{d}l}{r} \tag{2.23}$$

The potential at a point due to the most general charge distribution consisting of discrete charges as well as the three kinds of continuous charge distributions is given the grand sum of Eqs (2.19), (2.21), (2.22) and (2.23).

3. Potential Due to a Dipole

Two equal and opposite charges q and $-q$ separated by a distance a constitutes a dipole (see Sec. 1.12). The dipole moment **p** has a magnitude qa and points from the negative charge to the positive charge.

We will evaluate the potential at a point P at a distance r from the centre O of the dipole; the direction OP making an angle θ with the dipole moment **p** pointing, say, along the z-axis. From symmetry it is clear that the potential will not change as point P rotates about the z-axis, r and θ being fixed. Thus we need to find $\phi(r, \theta)$ for any plane containing this axis. Such a plane is shown in Fig. 2.4.

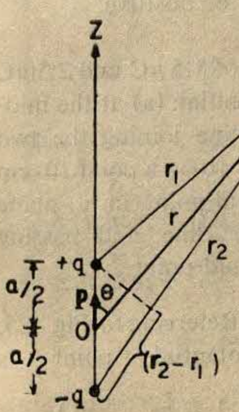


Fig. 2.4 *Potential at a point P in the field of a dipole*

Setting $q_1 = q$ and $q_2 = -q$ in Eq. (2.18) we get

$$\phi = \frac{q}{4\pi\epsilon_0}\left(\frac{1}{r_1} - \frac{1}{r_2}\right) = \frac{q}{4\pi\epsilon_0} \cdot \frac{r_2 - r_1}{r_1 r_2}$$

If point P is far away from the dipole, i.e. if $r_1, r_2 \gg a$, then it follows from Fig. 2.4 that

$$r_2 - r_1 \simeq a \cos\theta \qquad \text{and} \qquad r_1 r_2 \simeq r^2$$

and the potential reduces to

$$\phi = \frac{1}{4\pi\epsilon_0} \cdot \frac{p\cos\theta}{r^2} \tag{2.24}$$

Note that if $\theta = 90°$, $\phi = 0$, i.e., the potential at any point on the perpendicular bisector of a dipole is zero. This reflects the fact that no work is required to be done to bring in a test charge from infinity along the perpendicular bisector of a dipole. For a given r, ϕ has the greatest positive value for $\theta = 0$ and the greatest negative value for $\theta = 180°$.

Example 2.1 The electric potential at 10 cm from a point charge is +100 V. What is the magnitude and sign of the charge?

Solution: From Eq. (2.17) the magnitude of the charge is

$$q = 4\pi\epsilon_0 r\phi$$

Putting $\epsilon_0 = 8.85 \times 10^{-12}\ \text{C}^2\ \text{N}^{-1}\ \text{m}^{-2}$, $r = 10$ cm $= 0.10$ m and $\phi = 100$ V, we get

$$q = 1.1 \times 10^{-9}\ \text{C}$$

Since the potential is positive, the charge q must be positive.

Example 2.2 Two tiny spheres carrying charges of 1.5 μC and 2.5 μC are placed 30 cm apart. Find the electric potential: (a) at the mid-point of the line joining the two charges, and (b) at a point 10 cm from this mid-point in a plane normal to the line and passing through the mid-point.

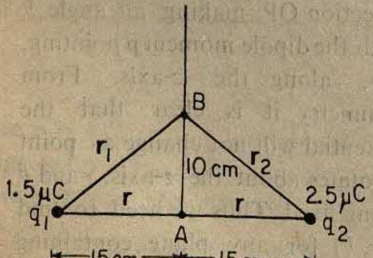


Fig. 2.5

Solution: (a) Referring to Fig. 2.5, the electric potential at point A is

$$\phi_A = \frac{1}{4\pi\epsilon_0}\left(\frac{q_1}{r} + \frac{q_2}{r}\right)$$

Substituting for ϵ_0, $q_1 = 1.5 \times 10^{-6}$ C, $q_2 = 2.5 \times 10^{-6}$ C and $r = 0.15$ m and solving we get

$$\phi_A = 2.4 \times 10^5 \text{ V}$$

(b) Now $r_1 = r_2 = \sqrt{10^2 + 15^2} = 18$ cm $= 0.18$ m

Therefore, the potential at B is

$$\phi_B = \frac{1}{4\pi\epsilon_0}\left(\frac{q_1}{r_1} + \frac{q_2}{r_2}\right) = 2.0 \times 10^5 \text{ V}$$

EXAMPLE 2.3 Four point charges $+q$, $+q$, $-q$, and $-q$ are placed respectively at the corners A, B, C and D of a square of side a. (a) Calculate the electric potential at O, the centre of the square. (b) If E and F are the mid-points of sides BC and CD respectively, what is the work that will be done in carrying an electron from: (i) O to E, and (ii) O to F? Given $q = 1.0 \times 10^{-6}$ C, $a = 0.10$ m and charge on an electron $= -1.6 \times 10^{-19}$ C.

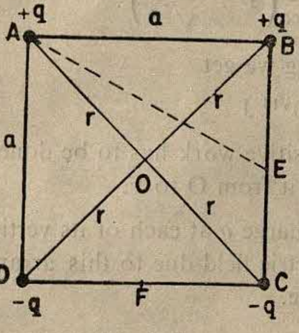


Fig. 2.6

Solution: (a) Referring to Fig. 2.6, the potential at O is given by

$$\phi_O = \frac{1}{4\pi\epsilon_0}\left(\frac{q}{\text{OA}} + \frac{q}{\text{OB}} - \frac{q}{\text{OC}} - \frac{q}{\text{OD}}\right)$$

$$= \frac{q}{4\pi\epsilon_0}\left(\frac{1}{r} + \frac{1}{r} - \frac{1}{r} - \frac{1}{r}\right) = 0$$

(b) Now $\quad (\text{AE})^2 = a^2 + \left(\frac{a}{2}\right)^2 = \frac{5a^2}{4}$

giving $\quad \text{AE} = \sqrt{5}\,\frac{a}{2}$

Similarly, $\quad \text{DE} = \sqrt{5}\,\frac{a}{2}$ and $\text{AF} = \text{BF} = \sqrt{5}\,\frac{a}{2}$

Potential at E is given by

$$\phi_E = \frac{1}{4\pi\epsilon_0}\left(\frac{q}{\text{AE}} + \frac{q}{\text{BE}} - \frac{q}{\text{DE}} - \frac{q}{\text{CE}}\right) = 0$$

Potential at F is given by

$$\phi_F = \frac{1}{4\pi\epsilon_0}\left(\frac{q}{\text{AF}} + \frac{q}{\text{BF}} - \frac{q}{\text{DF}} - \frac{q}{\text{CF}}\right)$$

Putting AF = BF = CF = DF = $\sqrt{5}\frac{a}{2}$ we get

$$\phi_F = \frac{q}{\pi\epsilon_0 a}\left(\frac{1}{\sqrt{5}} - 1\right) = -0.55\,\frac{q}{\pi\epsilon_0 a}$$

(i) Work done in carrying an electron of charge $-e$ from O to E is

$$W_{OE} = -e(\phi_E - \phi_O) = e(0 - 0) = 0$$

(ii) Work done in carrying a proton from O to F is

$$W_{OF} = -e(\phi_F - \phi_O) = -e\phi_F$$

$$= \frac{eq}{\pi\epsilon_0 a}\left(1 - \frac{1}{\sqrt{5}}\right)$$

Substituting for e, q, ϵ_0 and a and solving we get

$$W_{OF} = +3.2\times 10^{-14}\ \mathrm{J}$$

The potential ϕ_F is negative. Hence positive work has to be done on the negatively charged electron to carry it from O to F.

Example 2.4 A cube of side b has a charge q at each of its vertices. Determine the electric potential and electric field due to this arrangement of charges at the centre of the cube.

Solution: The distance of a vertex from the centre of the cube of side b is

$$r = \frac{\sqrt{3}\,b}{2}$$

Now the potential due to charge q at is the centre is $q/(4\pi\epsilon_0 r)$. Hence the potential due to the arrangement of eight charges (each of magnitude q) at the centre is

$$\phi = \frac{8q}{4\pi\epsilon_0 r} = \frac{4q}{\sqrt{3}\,\pi\epsilon_0 b}$$

We know that electric fields are to be added vectorially. From the symmetry of the eight charges with respect to the centre of the cube, it is evident the electric fields at the centre due to two opposite charges cancel in pairs (being equal and opposite). Hence the net electric field at the centre of the cube will be zero.

Example 2.5 Two point charges $-q$ and $+q$ are located at points $(0, 0, -a)$ and $(0, 0, a)$ respectively. (a) What is the electrostatic potential at points $(0, 0, z)$ and $(x, y, 0)$? (b) Obtain the dependence of potential on the distance r of the point from the origin when $r \gg a$. (c) How much work is done is moving a small test charge from the point $(5, 0, 0)$ to a point $(-7, 0, 0)$ along the x-axis? Does the answer change if the path of the test charge between the same points is not along the x-axis?

Solution:

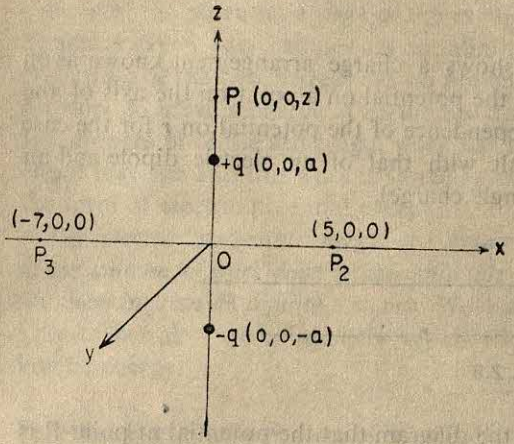


Fig. 2.7

(a) Referring to Fig. 2.7 the distance of point P₁ from charge $+q$ is $r_1 = z - a$ and from charge $-q$ is $r_2 = z + a$.

$$\therefore \quad \text{Potential at P}_1 = \frac{1}{4\pi\epsilon_0}\left(\frac{q}{r_1} - \frac{q}{r_2}\right) = \frac{q}{4\pi\epsilon_0} \cdot \frac{(r_2 - r_1)}{r_1 r_2}$$

$$= \frac{q}{4\pi\epsilon_0} \cdot \frac{2a}{(z^2 - a^2)} = \frac{p}{4\pi\epsilon_0(z^2 - a^2)} \qquad \text{(i)}$$

where $p = q(2a)$ is the dipole moment.

(b) If the distance of the point P₁ from the origin is r, then Eq. (i) reads

$$\phi\,(\text{at P}_1) = \frac{p}{4\pi\epsilon_0(r^2 - a^2)}$$

If $r \gg a$, this expression reduces to (neglecting a^2 compared to r^2)

$$\phi\,(\text{at P}_1) = \frac{p}{4\pi\epsilon_0} \cdot \frac{1}{r^2}$$

i.e. the potential at a far away point on the axis of a dipole varies as $1/r^2$.

(c) Any point on the perpendicular bisector passing through the centre of the dipole is at the same distance from the two charges. Hence the potential at point $P_2(5, 0, 0)$ and that at point $P_3(-7, 0, 0)$ are zero. Since P_2 and P_3 are at the same potential (zero), the potential difference between them is zero. Hence no work will be done in moving a charge from P_2 to P_3. The answer will not change if the path of the charge is changed because the work done is independent of the path taken.

EXAMPLE 2.6 Figure 2.8 shows a charge arrangement known as an 'electric quadrupole'. Find the potential on a point on the axis of the quadrupole. What is the dependence of the potential on r for the case $r \gg a$. Contrast your result with that of an electric dipole and an electric monopole (i.e. a single charge).

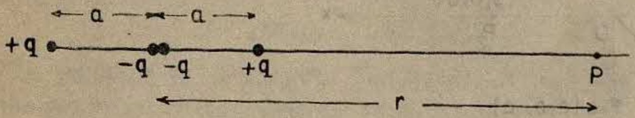


Fig. 2.8

Solution: It is clear from the diagram that the potential at point P is given by

$$\phi = \frac{1}{4\pi\epsilon_0}\left\{\frac{q}{(r-a)} - \frac{q}{r} - \frac{q}{r} + \frac{q}{(r+a)}\right\}$$

$$= \frac{2qa^2}{4\pi\epsilon_0 r(r^2-a^2)}$$

If $r \gg a$ this expression reduces to (dropping a^2 compared to r^2)

$$\phi = \frac{2qa^2}{4\pi\epsilon_0}\cdot\frac{1}{r^3} = \frac{Q}{4\pi\epsilon_0}\cdot\frac{1}{r^3}$$

where $Q = 2qa^2$ is called the electric quadrupole moment.

From Eqs (2.17) and Ex. 2.5 we notice that the potential varies as $1/r$ for a single charge (monopole), as $1/r^2$ for a dipole and as $1/r^3$ for a quadrupole.

2.4 ELECTRIC POTENTIAL ENERGY

If we raise a stone from the earth's surface, the work we have to do against the earth's gravitational attraction is stored as *potential energy* in the system comprising the stone and the earth. If the stone is released from the raised position, the potential energy of the system steadily changes into kinetic energy as the stone falls. When the stone comes to rest on the earth, the originally stored potential energy (which is completely converted into kinetic energy just before the stone hits the earth) is transformed into heat energy in the system of earth + stone.

A similar situation exists in electrostatics. When a stone rests on the earth, the potential energy of the earth-plus-stone system is zero. Similarly when two charges are an infinite distance apart, the system consisting of the two charges has no potential energy since it requires no work to move a charge at an infinite distance from another. If the two charges have the same sign, work has to be done to bring them closer together and this work is stored in the system of two charges in the form of electric potential energy.

The electric potential energy of a system of point charges is defined as the amount of work done to assemble this system of charges by bringing them in from an infinite distance. We assume that the charges were at rest when they were infinitely separated, i.e. they had no initial kinetic energy.

Potential Energy of a System of Two Point Charges

Suppose a point charge q_1 is at rest at a point P_1 in space. Suppose there is another point charge q_2 at rest at an infinite distance from q_1. How much work has to be done to bring the charge q_2 to a position P_2 at a distance r_{12} from q_1 (see Fig. 2.9)?

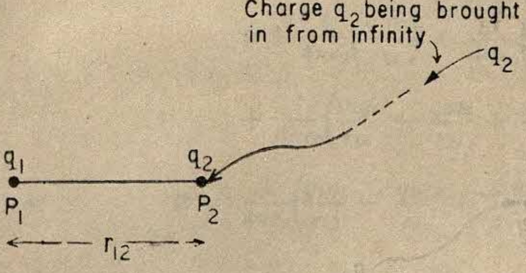


Fig. 2.9 *Potential energy of a system of two point charges*

We know that the electric potential at point P_2 caused by charge q_1 is given by

$$\phi\,(\text{at } P_2) = \frac{1}{4\pi\epsilon_0}\cdot\frac{q_1}{r_{12}} \tag{2.25}$$

If the charge q_2 is moved in from infinity to a position P_2, the work required is, from the definition of electric potential [Eq. (2.16)],

$$W = q_2\phi \text{ (at } P_2) \tag{2.26}$$

Combining these two equations and recalling that the work W is stored in the system $(q_1 + q_2)$ in the form of electric potential energy U, we have

$$U(=W) = \frac{1}{4\pi\epsilon_0} \cdot \frac{q_1 q_2}{r_{12}} \tag{2.27}$$

The potential energy in the system can be converted into kinetic energy. If the two charges are of the same kind $(q_1q_2 > 0)$ we could recover this energy by releasing them; they would accelerate away from each other on account of their repulsion in which case their potential energy would be converted into kinetic energy. If the charges were of the opposite kind $(q_1q_2 < 0)$, we could extract work from them by allowing them to come closer together; they would accelerate towards each other owing to their mutual attraction.

Potential Energy of a System of N Point Charges

We can easily generalize the result (2.27) and obtain an expression for the potential energy of a system of N point charges $q_1, q_2, q_3, \ldots, q_N$. The potential energy of this system is the total amount of work required to bring each charge in from infinity and assemble them at given positions. This can be done by bringing the charges in, one by one, from infinity (Fig. 2.10).

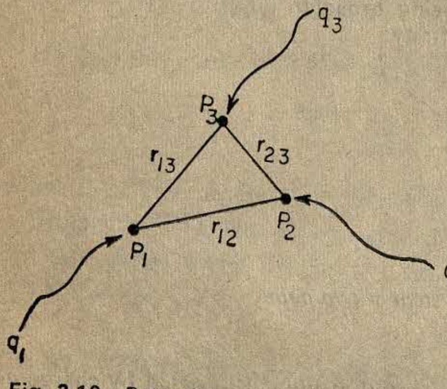


Fig. 2.10 *Potential energy of a system of N point charges*

It takes no work to bring in the first charge q_1, since there is no field yet to fight against (i.e. $W_1 = 0$). Now we bring it q_2 from infinity to a position P_2 at a distance r_{12} from q_1 at P_1. According to Eq. (2.26) the amount of work W_2 required is $q_2\phi_1$ (at P_2) where ϕ_1 is the potential due to q_1 at point P_2 where we are placing the charge q_2. Thus

$$W_2 = \frac{q_2}{4\pi\epsilon_0} \cdot \left(\frac{q_1}{r_{12}}\right)$$

where r_{12} is the distance between q_1 and q_2 when they are in position. Next we bring in q_3; this requires work $W_3 = q_3\phi_{1,2}$ (at P_3) where $\phi_{1,2}$ is the potential due to charges q_1 and q_2 at point P_3 where the charge q_3 is placed. Now, we know that

$$\phi_{1,2} = \frac{1}{4\pi\epsilon_0}\left(\frac{q_1}{r_{13}} + \frac{q_2}{r_{23}}\right)$$

Thus

$$W_3 = \frac{q_3}{4\pi\epsilon_0}\left(\frac{q_1}{r_{13}} + \frac{q_2}{r_{23}}\right)$$

Similarly the extra work to bring in q_4 will be

$$W_4 = \frac{q_4}{4\pi\epsilon_0}\left(\frac{q_1}{r_{14}} + \frac{q_2}{r_{24}} + \frac{q_3}{r_{34}}\right)$$

and so on till the Nth charge is brought in.

The total work necessary to assemble the first four charges is

$$\begin{aligned} W &= W_1 + W_2 + W_3 + W_4 \\ &= 0 + \frac{1}{4\pi\epsilon_0}\left(\frac{q_1q_2}{r_{12}}\right) + \frac{1}{4\pi\epsilon_0}\left(\frac{q_1q_3}{r_{13}} + \frac{q_2q_3}{r_{23}}\right) \\ &\quad + \frac{1}{4\pi\epsilon_0}\left(\frac{q_1q_4}{r_{14}} + \frac{q_2q_4}{r_{24}} + \frac{q_3q_4}{r_{34}}\right) \end{aligned}$$

or

$$W = \frac{1}{4\pi\epsilon_0}\left(\frac{q_1q_2}{r_{12}} + \frac{q_1q_3}{r_{13}} + \frac{q_1q_4}{r_{14}} + \frac{q_2q_3}{r_{23}} + \frac{q_2q_4}{r_{24}} + \frac{q_3q_4}{r_{34}}\right) \tag{2.28}$$

This expression for four charges gives us the general rule for N charges: take the product of each pair of charges, divide by their separation distance and add them up. Thus, the work done to assemble N point charges can be written as

$$W = \frac{1}{4\pi\epsilon_0} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j>i}}^{N} \frac{q_iq_j}{r_{ij}} \tag{2.29}$$

The restriction $j > i$ in Eq. (2.29) is to ensure that we do not count the same pair twice. A better way to ensure this is that we *intentionally* count each pair twice and then divide the result by 2. We should then write

$$W = \frac{1}{8\pi\epsilon_0} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{q_i q_j}{r_{ij}} \tag{2.30}$$

We must still avoid $j = i$. Notice that in this form the result does not depend on the *order* in which we assemble the charges, since every pair occurs in the sum. We can pull out q_i and write Eq. (2.30) as

$$W = \frac{1}{2} \sum_{i=1}^{N} q_i \left(\sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{1}{4\pi\epsilon_0} \cdot \frac{q_j}{r_{ij}} \right) \tag{2.31}$$

The term in round brackets is the potential ϕ_i at point P_i (the position of q_i) due to *all* other charges. Thus

$$W = \frac{1}{2} \sum_{i=1}^{N} q_i \phi_i \tag{2.32}$$

Potential Energy in a Fixed Electrostatic Field

Let us now consider a slightly different situation. How much work is necessary to a bring a charge, say q, to a particular position in space against the forces of a fixed electric field **E** caused by an arbitrary assembly of charges. We have seen earlier [Eq. (2.16)] that the work done to bring in a charge q, from infinity to a point in an electric field **E** where the potential is ϕ is given by

$$W = q\phi$$

This work W is stored as potential energy of the charge q in the field **E**. Thus

$$U = q\phi \tag{2.33}$$

We have also seen in Eq. (2.15) that the work done in moving a charge q from a point A to a point B in a given electric field is given by

$$W_{AB} = q(\phi_B - \phi_A)$$

Thus the change ΔU in potential energy as the charge q moves from A to B in an electrostatic field is the product of the charge q and the potential difference $(\phi_B - \phi_A)$. Since the electrostatic field is conserved, this change in potential energy is independent of the path which the charge takes to go from A to B, i.e,

$$\Delta U = U_B - U_A = q(\phi_B - \phi_A) \tag{2.34}$$

If q is *positive* and moves from a position of *lower* potential to one of *higher* potential, its own potential energy *increases*. If q is *negative*, its potential energy *increases* if it moves from a *higher* to a *lower* potential.

The Electron-Volt

The SI unit of potential energy is the *joule*. In atomic physics and chemistry, a more convenient unit called the *electron-volt* (written as eV) is used. *An electron-volt is the potential energy gained or lost by an electron in moving through a potential difference of 1 volt.* Since the magnitude of charge on an electron is 1.6×10^{-19} C,

$$1 \text{ eV} = 1.6 \times 10^{-19} \text{ J}$$

We have found the potential energy of a single charge when it is placed at a point in an externally produced electric field. What happens if two charges q_1 and q_2 are placed at points A and B in an electrostatic field produced by external sources as shown in Fig. 2.11?

The two charges will have their own potential energies U_A and U_B due to the interaction of each with the field [Eq. (2.32)] in addition to their mutual potential energy U_{AB} [Eq. (2.27)]. Thus the total potential energy of the two charges is given by

$$\begin{aligned} U &= U_A + U_B + U_{AB} \\ &= q_1\phi_A + q_2\phi_B + \frac{1}{4\pi\epsilon_0} \cdot \frac{q_1 q_2}{r_{12}} \qquad (2.35) \end{aligned}$$

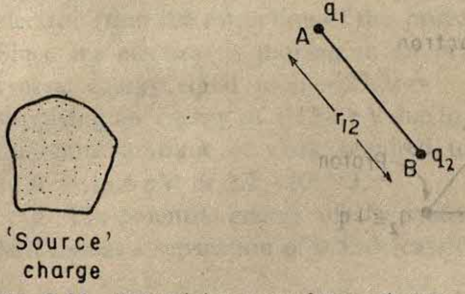


Fig. 2.11 *Potential energy of two charges in an external electrostatic field*

EXAMPLE 2.7 Two positive point charges of 12 μC and 8 μC are 10 cm apart. Find the work done in bringing them 4 cm closer.

Solution: $q_1 = 12 \times 10^{-6}$ C, $q_2 = 8 \times 10^{-6}$ C

Initial separation (r_i) = 10 cm = 0.10 m

$$\text{Final separation } (r_f) = 6 \text{ cm} = 0.06 \text{ m}$$

$$\text{Initial potential energy } (U_i) = \frac{1}{4\pi\epsilon_0} \cdot \frac{q_1q_2}{r_i}$$

$$\text{Final potential energy } (U_f) = \frac{1}{4\pi\epsilon_0} \frac{q_1q_2}{r_f}$$

∴ Change in potential energy is

$$U_f - U_i = \frac{q_1q_2}{4\pi\epsilon_0}\left(\frac{1}{r_f} - \frac{1}{r_i}\right) = \frac{q_1q_2}{4\pi\epsilon_0} \cdot \frac{(r_i - r_f)}{r_ir_f}$$

Now work done = change in potential energy. Substituting the values of q_1, q_2, r_i, r_f and ϵ_0 and solving, we get

$$\text{Work done} = 5.8 \text{ J}$$

EXAMPLE 2.8 A neutral hydrogen molecule has two protons and two electrons. If one of the electrons is removed we get a hydrogen molecular ion (H_2^+). In the ground state of H_2^+ the two protons are separated by roughly 1.5Å and the electron is roughly 1Å from each proton. Estimate the potential energy of the system.

Solution: The total potential energy of the arrangement of charges (Fig. 2.12) is the sum of the energies of each pair of charges. From Eq. (2.28) the potential energy of the system comprising the three charges q_1, q_2 and q_3 is

$$\begin{aligned} U &= W_1 + W_2 + W_3 \\ &= \frac{1}{4\pi\epsilon_0}\left(\frac{q_1q_2}{r_{12}} + \frac{q_1q_3}{r_{13}} + \frac{q_2q_3}{r_{23}}\right) \qquad \text{(i)} \end{aligned}$$

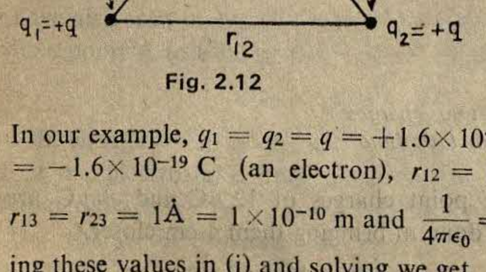


Fig. 2.12

In our example, $q_1 = q_2 = q = +1.6\times 10^{-19}$ C (two protons) $q_3 = -q = -1.6\times 10^{-19}$ C (an electron), $r_{12} = 1.5\text{Å} = 1.5\times 10^{-10}$ m, and $r_{13} = r_{23} = 1\text{Å} = 1\times 10^{-10}$ m and $\frac{1}{4\pi\epsilon_0} = 9\times 10^9 \text{ Nm}^2\text{C}^{-2}$. Substituting these values in (i) and solving we get

$$\begin{aligned} U &= -30.7\times 10^{-19} \text{ J} \\ &= -19.2 \text{ eV} \end{aligned}$$

The fact that the energy is negative means that so much energy is needed to dismantle the system, i.e. to remove the charges to an infinite separation from each other.

EXAMPLE 2.9 In a hydrogen atom, the electron and proton are bound together at a separation of about 0.53Å. (a) Estimate the potential energy of the system in electron volts, taking the zero of potential energy at an infinite separation of the electron from the proton. (b) What is the minimum work required to free the electron from the proton, given that the kinetic energy of the electron in its orbit is half the potential energy in (a)? (c) What is the answer to (a) and (b) above if the zero of potential energy is taken at 1.06Å separation?

Solution

Charge on an electron $(q_1) = -1.6 \times 10^{-19}$ C
Charge on a proton $(q_2) = +1.6 \times 10^{-19}$ C
Separation $(r_{12}) = 0.53\text{Å} = 0.53 \times 10^{-10}$ m

(a) If the zero of potential energy is taken to be at infinite separation, the potential energy of the electron-proton system is

$$U = \frac{1}{4\pi\epsilon_0} \cdot \frac{q_1 q_2}{r_{12}}$$

$$= -4.35 \times 10^{-18}\ \text{J} \qquad = -27.2\ \text{eV}$$

(b) If the electron were at rest, 27.2 eV of energy will have to be supplied (or 4.35×10^{-18} J of work will have to be done) to free the electron from the attraction of the proton and remove it to infinity. Since the electron is moving in an orbit (round the proton) with a kinetic energy equal to $\frac{1}{2}(-27.2) = -13.6$ eV, the electron itself is supplying an energy of $+13.6$ e V due to centrifugal action. Hence the minimum amount of work required to free the electron $= 27.2 - 13.6 = 13.6$ eV or 2.2×10^{-18} J.

(c) The potential energy of the system at a separation of 1.06Å = half that at a separation of 0.53Å [case (a) above]

$$= -13.6\ \text{eV}$$

If the energy at a separation of 1.06Å is taken to be zero (instead of -13.6 eV), the potential energy in case (a) would obviously be $-27.2 - (-13.6) = -13.6$ eV. In case (b) this much energy is being supplied by the electron itself by virtue of its orbital motion round the proton. Hence the minimum work to pull the electron from the proton will be zero. So if the zero of potential energy is taken at a separation of 1.06Å, the total energy of the hydrogen atom is zero.

2.5 FROM ELECTRIC POTENTIAL TO ELECTRIC FIELD

We know that the potential difference between to points A and B in an electrostatic field **E** is given by

$$\phi_B - \phi_A = -\int_A^B \mathbf{E}\cdot d\mathbf{l} \tag{2.36}$$

Hence if **E** is given in a region, then we can find the potential difference between any two points in that region (and the potential ϕ at any point) by simply evaluating the line integral of the field. Is the reverse also true? Can we find the electric field **E** in a region if the potential ϕ is known at all points in that region? The answer is yes. To show this we will, for simplicity, consider a one-dimensional situation in which both potential and electric field depend only on the x-coordinate of a point and the field has only the x-component E_x (Fig. 2.13).

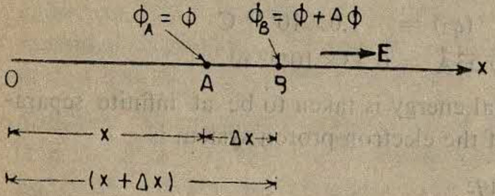


Fig. 2.13 *Relation between potential and field*

Suppose A and B are two neighbouring points at distances x and $(x + \Delta x)$ from the origin O so that the distance Δx between them is small. Let the potential ϕ_A at A be ϕ and the potential ϕ_B at B be $(\phi + \Delta\phi)$ so that the potential difference between A and B is

$$\phi_B - \phi_A = (\phi + \Delta\phi) - \phi = \Delta\phi$$

The work done in carrying a test charge q_0 (small and positive) from B to A is given by [see Eq. 2.15)]

$$W = q_0\Delta\phi \tag{2.37}$$

We can calculate this work from another point of view also. The electric field E_x exerts a force q_0E_x in the x-direction. Hence to carry the charge from B to A, we have to apply a force F equal and opposite to $q_0 E_x$, i.e. the force to applied is

$$F = -q_0E_x$$

Assuming that the force F remains almost constant as we move the charge from B to A (a very small distance Δx), the work done will be

$$W = \text{force} \times \text{distance} = F \times \Delta_x = -q_0E_x \cdot \Delta x \tag{2.38}$$

Equating the two expressions (2.37) and (2.38) for W we have

$$q_0 \Delta\phi = -q_0 E_x \cdot \Delta x$$

or

$$E_x \Delta x = -\Delta\phi$$

Dividing both sides by Δx we get

$$E_x = -\frac{\Delta\phi}{\Delta x}$$

Taking the limit as Δx tends to zero gives

$$E_x = -\frac{\partial\phi}{\partial x} \tag{2.39}$$

The quantity $\partial\phi/\partial x$ is the rate at which the potential changes with distance, this quantity is called *potential gradient*. Equation (2.39) shows that *the electric field is equal to the negative of the potential gradient*.

Alternative Proof We can prove relation (2.39) very simply if we use the definition of potential difference in terms of the line integral of the electric field [Eq. (2.36)]. If we connect the two points A (at x) and B (at $x + \Delta x$) by a short straightline segment, then $\mathrm{d}l = \Delta x$ and the line integral in Eq. (2.36) is just a single term $-E_x \Delta x$ and it reads

$$\Delta\phi = -E_x \Delta x$$

Dividing by Δx and taking the limit $\Delta x \to 0$ immediately yields relation (2.39). For a general proof of relation (2.39) in three dimensions, refer to Supplement S1 at the end of this chapter.

Relation (2.39) gives an alternative SI unit for electric field **E**. The unit for ϕ is the volt and that for x is the metre, and hence **E** can also be expressed in unit *volts per metre* ($\mathrm{Vm^{-1}}$). We have so far used newtons per coulomb ($\mathrm{NC^{-1}}$) as the unit of **E**. The two units are equivalent since

$$\text{volt} = \frac{\text{joule}}{\text{coulomb}}$$

$$= \frac{\text{newton metre}}{\text{coulomb}} \qquad (\because 1\,\mathrm{J} = 1\,\mathrm{Nm})$$

or

$$\frac{\text{volt}}{\text{metre}} = \frac{\text{newton}}{\text{coulomb}}$$

or

$$1\,\mathrm{Vm^{-1}} = 1\,\mathrm{NC^{-1}}$$

Relation (2.39) shows how we can find electric field **E** if potential ϕ is known. It is a major simplification because potential (being a scalar) is much easier to work with than electric field which is a vector; we are saved from carrying out cumbersome vector operations. The following simple example illustrates how **E** can be found if ϕ is given.

EXAMPLE 2.10 Electric potential at a point at a distance r from a point charge q is given by

$$\phi = \frac{1}{4\pi\epsilon_0} \cdot \frac{q}{r}$$

Find electric field **E** at that point.

Solution: From symmetry, **E** must be directed radially outward from a (positive) point charge. Hence the relevant derivative of ϕ is $\partial\phi/\partial r$ instead of $\partial\phi/\partial x$. Thus the relation between E and ϕ is

$$E = -\frac{\partial\phi}{\partial r} = -\frac{\partial}{\partial r}\left(\frac{1}{4\pi\epsilon_0} \cdot \frac{q}{r}\right)$$

$$= -\frac{q}{4\pi\epsilon_0} \frac{\partial}{\partial r}\left(\frac{1}{r}\right)$$

$$= \frac{q}{4\pi\epsilon_0} \cdot \frac{1}{r^2}$$

This result agrees exactly with our earlier result (see Chapter 1).

EXAMPLE 2.11 Two large parallel metal plates A and B are separated by a distance $l = 2.5$ cm (see Fig. 2.14). The plate B has a potential of 2500 V relative to plate A. An electron is liberated from a point near the middle of plate A. How long does it take to reach plate B. Charge on an electron $(e) = -1.6 \times 10^{-19}$ C and mass of an electron $= 9.1 \times 10^{-31}$ kg.

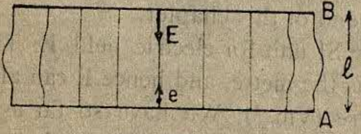


Fig. 2.14

Solution: Between two large parallel plates, the electric field is uniform except near the edges as shown in Fig. 2.14. Hence, near the middle the potential gradient is uniform and has a value ϕ/l, i.e.

$$\frac{\partial\phi}{\partial x} = \frac{\phi}{l} = \frac{2500\ \text{V}}{2.5 \times 10^{-2}\ \text{m}} = 1.0 \times 10^5\ \text{Vm}^{-1}$$

$$\therefore \quad \text{Electric field } (E) = -\frac{\partial\phi}{\partial x} = -1.0 \times 10^5\ \text{Vm}^{-1}$$

Hence the force on an electron of charge e is

$$F = eE = (-1.6\times 10^{-19})\times(-1.0\times 10^{5})$$
$$= 1.6\times 10^{-14}\,\mathrm{N}$$

$$\therefore \quad \text{Acceleration } a = \frac{F}{m} = \frac{1.6\times 10^{-14}}{9.1\times 10^{-31}} = 1.76\times 10^{16}\ \mathrm{ms^{-2}}$$

which is very large compared to the acceleration due to gravity, $g = 9.8\ \mathrm{ms^{-2}}$. Since the electron is at rest when librated, the time it takes to traverse the distance l between the plates is given by

$$t = \sqrt{\frac{2l}{a}}$$

Substituting the values of l and a we get

$$t = 1.7\times 10^{-9}\ \mathrm{s}$$

Equipotential Surfaces

Any two-dimensional surface over which the electric potential is constant is called an equipotential surface. The surface may be a surface of a body or simply a surface in space. For example, as we will see later, the surface of a conductor is an equipotential surface. Equipotential surfaces can be drawn throughout any space in which there is an electric field as we shall now explain.

Let us consider the simplest case of the electric field of an isolated point charge q. At a distance r from the charge, the potential is

$$\phi = \frac{q}{4\pi\epsilon_0 r}$$

A sphere of radius r with centre at q is, therefore, an equipotential surface of potential $q/(4\pi\epsilon_0 r)$. In fact all spheres centred on q are equipotential surfaces, whose potentials are inversely proportional to r (Fig. 2.15).

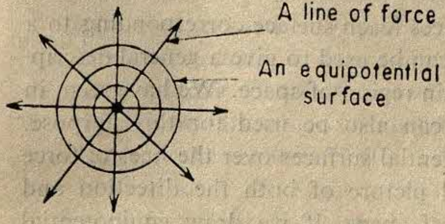


Fig. 2.15 *Equipotential surfaces and lines of force around a positive point charge*

An important property of an equipotential surface is that, along any direction lying on the surface, there is no electric field, because there is no potential gradient [see Eq. (2.39)]. This means that *the electric field* **E** (*and hence the lines of force*) *are always at right angles to an equipotential surface.* If **E** were not at right angles to the surface, it would have a component lying in that surface which, by definition, is not possible. Also *no work is required to move a charge along an equipotential surface* as is obvious from Eq. (2.15) since on an equipotential surface, the phase difference ($\phi_B - \phi_A$) between any two points is zero.

The fact that **E** is perpendicular to the equipotential surface is obvious in the case of a single point charge; **E** is radial and the surfaces are concentric spheres (Fig. 2.15). For more general fields (such as those due to two or more point charges or uniform charge distributions) the equipotential surfaces are more complicated. On any equipotential surface, the electric field **E** at any point is perpendicular to the surface. This can also be argued as follows:

Take a point P and draw an equipotential surface through it (Fig. 2.16). Since, by definition, the potential difference between any two points on the surface is zero, it follows from Eq. (2.12) that the line integral of **E** along the surface is zero, i.e. $\mathbf{E}\cdot d\mathbf{l} = 0$ where dl is a line element on the surface. Now $\mathbf{E}\cdot d\mathbf{l} = E\,dl\cos\theta = 0$, therefore $\theta = 90^\circ$ i.e. **E** and **dl** are at right angles to each other as shown in Fig. 2.16.

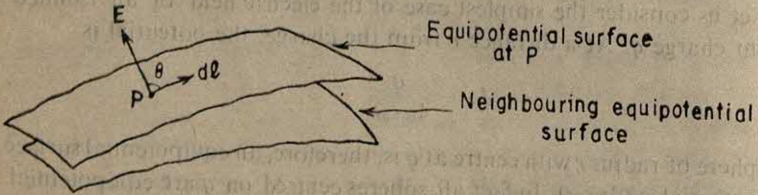

E
θ
dℓ
P
Equipotential surface at P
Neighbouring equipotential surface

Fig. 2.16 *Relation between electric field and equipotential surface*

A family of equipotential surfaces (each surface corresponding to a different value of the potential) can be used to give a general description of the electric field in a certain region of space. We have seen, in Chapter 1, that the lines of force can also be used for this purpose. The advantage of using equipotential surfaces over the lines of force is that such surfaces give a visual picture of both the direction and the magnitude of **E** in a region of space. If we draw equipotential surfaces at regularly spaced intervals of ϕ, we can distinguish regions of strong field from those of weak field: the surfaces are closer together where the field is stronger and farther apart where the field is weaker. This follows from Eq. (2.39), due to the negative sign, **E** *is in the direction of the steepest or fastest decrease in* ϕ. In addition the direction of **E** is perpendicular at each point on the surface.

2.6 GAUSS'S THEOREM FOR ELECTROSTATICS

We have obtained many useful results from the study of line integrals of electrostatic fields. This study has helped us to define electric potential, electric potential energy, equipotential surfaces, etc. We have learnt how to compute the field of a charge distribution, the force on a charge placed in this field and how to obtain electric field from the potential. In principle this is all that we have to learn in the subject of electrostatics. Unfortunately, the integrals involved in computing **E** can be very formidable even for reasonably simple charge distributions. Much of the rest of electrostatics is devoted to assembling tools and tricks for avoiding these integrals. One such extremely useful tool is the *Gauss's theorem* which exploits the properties of surface integrals of electrostatic fields. We will find that these integrals contain information about the *sources* of the field.

Because of its usefulness (as we will presently see) we will state Gauss's theorem at the very outset deferring its proof for a while. Let *S* be any closed surface in space enclosing a volume *V*. Then the theorem states *that the surface integral over S of the electrostatic field* **E** *due to any sources is proportional to the total charge q contained inside S in the volume V, i.e*

$$\int_S \mathbf{E}\cdot d\mathbf{S} = \frac{q}{\epsilon_0} \tag{2.40}$$

where dS is an element of area; the direction of each area element dS of *S* is along the outward drawn normal to the surface at every point. At each point on *S* the electric field **E** will have contributions from charges within *V* and charges outside. The theorem states that *the charges outside V do not contribute to the surface integral; only the charges inside V are all that matter.* Any closed surface that is chosen to calculate the surface integral is called a *Gaussian surface.* Note that *q* in the theorem (2.40) is the *net* charge (taking account of the signs) inside *S*, the exact location of *q* inside *S* does not affect the value of the surface integral (Refer to Appendix I at the end of the book for the meaning of 'surface integral'.)

Proof of Gauss's Theorem

We will verify the theorem for the simplest case. Suppose the field **E** is that of a single isolated positive point charge q and the surface S is a sphere of radius r centred on the point charge (Fig. 2.17). What is the value of the surface integral of **E** over S? The answer is easy to find, as we show below.

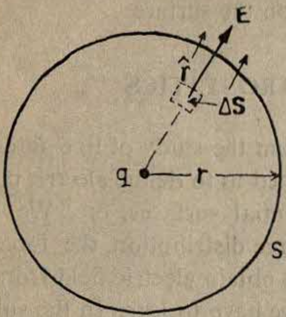


Fig. 2.17 *Evaluation of surface integral of electric field due to a point charge*

The magnitude of **E** at every point on the surface S is the same $(= q/4\pi\epsilon_0 r^2)$ and its direction is the same as that of the outward drawn normal (if q is positive) at that point. Any element of area dS on S points radially outwards parallel to **E** (it will be antiparallel to **E** if q is negative) and has a magnitude dS. If $\hat{\mathbf{r}}$ is a unit vector from q to the surface element, then:

$$\mathbf{E} = \frac{q}{4\pi\epsilon_0} \cdot \frac{\hat{\mathbf{r}}}{r^2}$$

and
$$\mathrm{d}\mathbf{S} = \Delta S \cdot \hat{\mathbf{r}}$$

The dot product of **E** and dS is

$$\mathbf{E} \cdot \mathrm{d}\mathbf{S} = \frac{q}{4\pi\epsilon_0} \cdot \frac{\Delta S}{r^2} (\hat{\mathbf{r}} \cdot \hat{\mathbf{r}})$$

$$= \frac{q}{4\pi\epsilon_0} \cdot \frac{\Delta S}{r^2} \quad (\because \hat{\mathbf{r}} \cdot \hat{\mathbf{r}} = 1)$$

Integrating over S and recalling that $\int_S \Delta S =$ total area of the sphere of radius $r = 4\pi r^2$, we get

$$\int_S \mathbf{E} \cdot \mathrm{d}S = \frac{q}{4\pi\epsilon_0} \frac{1}{r^2} (4\pi r^2)$$

$$= \frac{q}{\epsilon_0}$$

which verifies Gauss's theorem. Notice that the radius of the sphere cancels out, for while the surface area *increases* as r^2, the electric field *decreases* as $1/r^2$ and the product is independent of r. Thus we could choose a sphere of any radius, the surface integral would be independent of the radius.

It may be mentioned that we choose a sphere because it is the most convenient choice; Gauss's theorem holds for any general surface. For a proof of this theorem for a general surface, refer to Supplement *S*2 at the end of this chapter.

Generalization to a System of N Point Charges

Now suppose that instead of a single charge, we have a bunch of N charges all scattered about within the surface. According to the principle of superposition, the total field **E** is simply the vector sum of all the individual fields due to individual charges. Thus

$$\mathbf{E} = \sum_{n=1}^{N} \mathbf{E}_n$$

The surface integral of **E** over S then reads

$$\int_S \mathbf{E}\cdot d\mathbf{S} = \sum_{n=1}^{N} \int_S \mathbf{E}_n \cdot d\mathbf{S}$$

$$= \frac{1}{\epsilon_0} \sum_{n=1}^{N} q_n = \frac{Q}{\epsilon_0} \tag{2.41}$$

where $Q = \sum_{n=1}^{N} q_n$ is the total charge inside the surface. The total charge is found by taking the sign of each charge q_n in the sum.

Some Applications of Gauss's Theorem

Gauss's theorem can be used to evaluate **E** due to a charge distribution; one takes a convenient Gaussian surface to evaluate the surface integral. Conversely, if **E** is known for all points on a closed surface, Gauss's theorem can be used to compute the charge inside. If **E** has an *outward* component for every point of the closed surface, then the surface integral will be positive indicating that the net charge inside the surface is *positive*. If **E** has an *inward* component for every point on the closed surface, there must be a net *negative* charge within the surface. If the surface encloses no charge and the field **E** is due to a charge outside the surface, then, according to Gauss's theorem, the surface integral will be zero. This is because of the fact that the lines of force originating from a charge outside the surface, directly pass through the surface; the contribution to the integral on one side cancelling that on the other.

Let us now consider a few simple applications of Gauss's theorem.

In each example, a judicious choice of Gaussian surface will be made depending on the symmetry of the electric field. Even though Gauss's theorem holds for any surface (sphere, cube, cylinder, etc), we will choose the most convenient Gaussian surface so that the surface integral is easily evaluated.

(a) *Electric Field of a Line of Charge* Consider a thin infinitely long straight rod with a uniform linear charge density λ Cm^{-1}. What is the expression for electric field **E** at a distance r from the rod (assumed to be just a line)? Figure 2.18 shows a section of the rod.

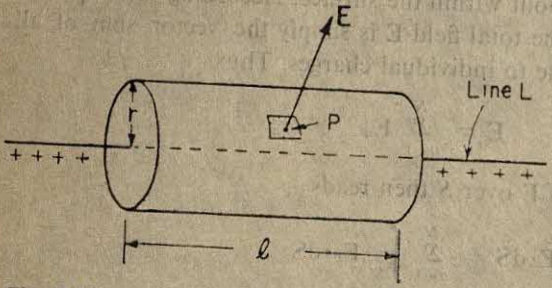


Fig. 2.18 *A straight line of charge, showing a cylindrical Gaussian surface*

From symmetry, **E** due to a uniform linear charge can only be radially directed. At any point P at a perpendicular distance r from the line, it is easy to see that **E** at point P will lie in the plane containing P and its direction will be radially outward (if the charge is positive) or inward (if the charge is negative), the magnitude of **E** depending only on r. This suggests that we should choose the Gaussian surface to be a circular cylinder of radius r and length, say, l, closed at each end by plane caps normal to the axis of the line L (Fig. 2.18).

Now **E** is constant over the cylindrical surface. The surface integral over the cylindrical surface is simply $E\times(2\pi rl)$ where E is the magnitude of electric field **E** on the surface and $(2\pi rl)$ is the area of the cylindrical surface. The surface integral over the end caps is zero because **E** here lies along the surface at every point.

The charge enclosed inside the Gaussian surface is

$$q = \lambda l$$

From Gauss's theorem, we have

$$\text{Surface integral} = \frac{q}{\epsilon_0}$$

or $$E \times (2\pi r l) = \frac{q}{\epsilon_0}$$

or $$E = \frac{\lambda}{2\pi\epsilon_0 r} \tag{2.42}$$

This expression gives the magnitude of electric field **E** at a distance r from the line. The direction of **E** is radially outward if the line charge is positive, it is radially inward if the line charge is negative.

(b) *Electric Field due to a Thin Sheet of Charge* Consider a thin flat sheet, infinite in extent, with a uniform surface charge density σ (measured in Cm^{-2}). We will use Gauss's theorem to evaluate **E** at a distance r in front of the sheet (Fig. 2.19).

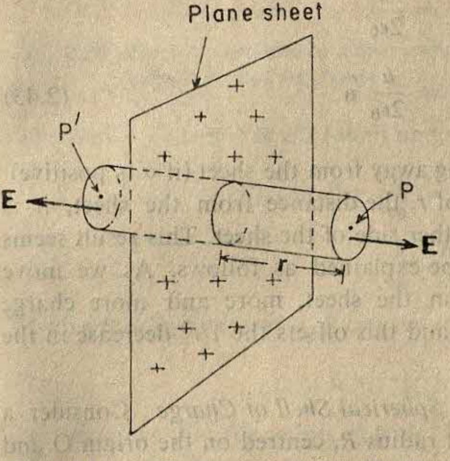


Fig. 2.19 *Using Gauss's theorem to find the field due to an infinite flat sheet of charge*

From the symmetry of the two sides of the sheet, we see that on both sides **E** points outwards perpendicular to the sheet (if σ is positive) and inwards (if σ is negative). Also from symmetry, the field must have the same magnitude and the opposite direction at two points P and P′ equidistant from the sheet on opposite sides. These facts suggest a convenient Gaussian surface; it is a cylinder of cross-sectional area, say A, extending up to P and P′ on two sides of the sheet (where **E** is to be evaluated); the axis of the cylinder being perpendicular to the sheet (Fig. 2.19).

Since **E** is perpendicular to the plane end-faces of the cylinder, these two faces contribute $(EA + EA) = 2EA$ to the surface integral whereas the curved sides of the cylinder contribute nothing, where E is the magnitude of **E**. Thus

$$\int_S \mathbf{E}\cdot d\mathbf{S} = 2\,EA$$

From Gauss's theorem

$$\int_S \mathbf{E}\cdot d\mathbf{S} = \frac{q}{\epsilon_0}$$

where $q = \sigma A$ is the enclosed charge. Hence

$$2EA = \frac{\sigma A}{\epsilon_0}$$

or
$$E = \frac{\sigma}{2\epsilon_0}$$

and
$$\mathbf{E} = \frac{\sigma}{2\epsilon_0}\,\hat{\mathbf{n}} \tag{2.43}$$

where $\hat{\mathbf{n}}$ is a unit vector pointing away from the sheet (if σ is positive).

Note that E is independent of r, the distance from the sheet, it is the same at all points on either side of the sheet. This result seems surprising at first sight. It can be explained as follows. As we move further and further away from the sheet, more and more charge comes into our 'field of view' and this offsets the $1/r^2$ decrease in the field from any particular piece.

(c) *Electric Field due to a Thin Spherical Shell of Charge* Consider a thin spherical shell of charge of radius R, centred on the origin O and having a uniform surface charge density σ (in $\mathrm{Cm^{-2}}$). It follows that the field is purely radial because of the spherical symmetry of the field. Let us calculate the magnitude E of the radial electric field at a point P at a distance r from the centre O of the shell. We can apply Gauss's theorem to the two possible cases; (a) P lying inside the shell ($r < R$) and (b) outside the shell ($r > R$).

From points (such as P′) inside the shell, the Gaussian surface is a sphere S' containing P′ centred on O. Since this sphere does not enclose any charge (the shell is outside the sphere) if follows from Gauss's theorem that **E** is zero everywhere inside the shell ($r' < R$).

For points (such as P) outside the shell the Gaussian surface S is a sphere containing P centred on O. The total charge q inside this sphere is the charge on the shell of radius R and area $4\pi R^2$.

$$q = 4\pi R^2\sigma$$

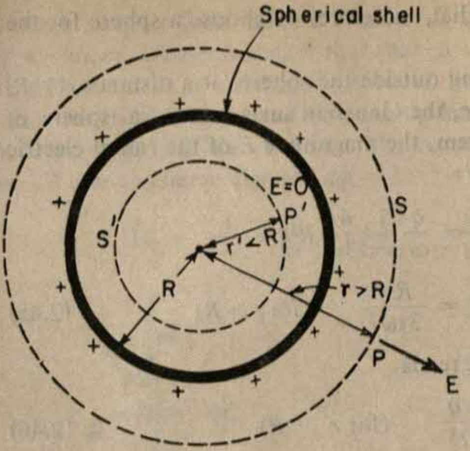


Fig. 2.20 *Electric field due to a thin uniform spherical shell of charge*

The surface integral is $E\times(4\pi r^2)$ and Gauss's theorem gives

$$E\times(4\pi r^2) = \frac{q}{\epsilon_0} = \frac{4\pi R^2\sigma}{\epsilon_0}$$

or
$$E = \frac{\sigma}{\epsilon_0}\cdot\frac{R^2}{r^2} \tag{2.44}$$

Notice that in terms of q, E can be written as

$$E = \frac{1}{4\pi\epsilon_0}\cdot\frac{q}{r^2}$$

which shows that for points outside, the shell behaves as though the charge were concentrated at the centre of the shell. The direction of the field is along the outward normal (if q is positive) or inward normal (if q is negative) to the surface of the shell at each point.

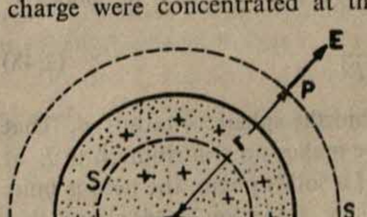


Fig. 2.21 *Electric field due to a solid sphere of charge*

(*d*) *Electric Field Due to a Solid Sphere of Charge* Figure 2.21 shows a solid sphere of radius R and having a uniform volume charge density ρ (measured in $\mathrm{Cm^{-3}}$). The total charge q on the sphere is

$$q = \frac{4\pi}{3}R^3\rho$$

The electric field is again radial, therefore, we choose a sphere for the Gaussian surface (Fig. 2.21).

For points (such as P) lying outside the sphere, at a distance $r(>R)$ from the centre of the sphere, the Gaussian surface S is a sphere of radius r. From Gauss's theorem, the magnitude E of the radial electric field at this point is given by

$$4\pi r^2 E = \frac{q}{\epsilon_0} = \frac{4\pi}{3\epsilon_0} R^3 \rho$$

$$\text{or} \quad E = \frac{R^3\rho}{3\epsilon_0 r^2} \qquad (\text{for } r > R) \tag{2.45}$$

In terms of total charge q, it reads

$$E = \frac{1}{4\pi\epsilon_0} \cdot \frac{q}{r^2} \qquad (\text{for } r > R) \tag{2.46}$$

For points (such as P′) inside the sphere at a distance r' $(< R)$ from the centre of the sphere, the Gaussian surface S' is a sphere of radius r'. The total charge q' enclosed by this surface is

$$q' = \frac{4\pi}{3} r'^3 \rho$$

From Gauss's theorem, the electric field E' at P′ is given by

$$4\pi r'^2 E' = \frac{q'}{\epsilon_0} = \frac{4\pi r'^3 \rho}{3\epsilon_0}$$

$$\text{or} \qquad E' = \frac{r'\rho}{3\epsilon_0} \qquad (\text{for } r' < R) \tag{2.47}$$

In terms of total charge q of the sphere, it reads (since $\rho = 3q/4\pi R^3$)

$$E' = \frac{q}{4\pi\epsilon_0} \cdot \frac{r'}{R^3} \tag{2.48}$$

Here q' is the part of q contained within the sphere of radius r'. That part of q which lies outside this sphere makes no contribution to E at distances $r' < R$. Thus in the case of a solid sphere, the field is finite at all points inside and outside the sphere. Equation (2.46) shows that for points outside the sphere it behaves as though all the charge were concentrated at its centre. Note that the sphere cannot be a conductor because, as we will see, the charge will reside on its surface.

EXAMPLE 2.12 Obtain the expression (2.42) for the electric field due to a long thin rod of uniform linear charge density λ using Coulomb's law instead of Gauss's theorem.

Solution: Figure 2.22 shows a section of an infinite line of charge whose linear charge density that has a constant value λ. We will calculate the field at a distance r from the line. We divide the line into a large number of very small elements of length dx and find the sum of the contributions of the elements. The magnitude of the field contribution dE due to charge element $dq(=\lambda dx)$ is given by Coulomb's law:

$$\mathrm{d}E = \frac{1}{4\pi\epsilon_0}\cdot\frac{dq}{y^2} = \frac{1}{4\pi\epsilon_0}\cdot\frac{\lambda dx}{(r^2+x^2)} \quad \text{(i)}$$

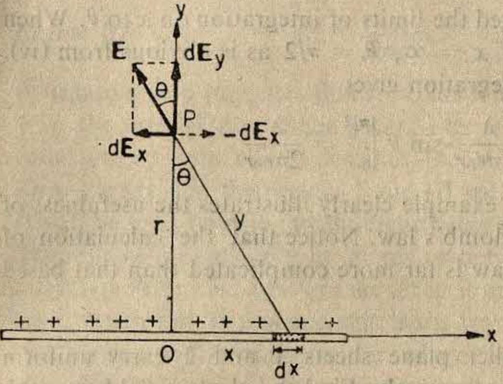


Fig. 2.22 *Electric field due to an infinite line of charge*

The vector $\mathbf{dE}$ in Fig. 2.22 has the two rectangular components $\mathrm{d}E_x$ and $\mathrm{d}E_y$. The x-component of the total field $\mathbf{E}$ must be zero because every charge element on the right of O has a corresponding element to the left of O such that their field contributions in the x directions cancel. Thus $\mathbf{E}$ points entirely along the y direction. It is clear from the diagram that

$$\mathrm{d}E_y = \mathrm{d}E\cos\theta$$

Therefore, the field E due to the entire line of charge is

$$E = E_y = \int_{x=-\infty}^{x=+\infty}\cos\theta\,\mathrm{d}E$$

$$= 2\int_{x=0}^{x=\infty}\cos\theta\,\mathrm{d}E \quad \text{(ii)}$$

Note that we have changed the lower limit of integration and introduced a factor of 2. The reason is that the contributions to $\mathrm{d}E_y$ from the right and left halves of the rod are equal. Using (i) in (ii) we get

$$E = \frac{\lambda}{2\pi\epsilon_0}\int_{x=0}^{x=\infty}\cos\theta\,\frac{\mathrm{d}x}{(r^2+x^2)} \quad \text{(iii)}$$

From Fig. 2.22 the relation between θ and x is

$$x = r \tan \theta \tag{iv}$$

Differentiating, we get

$$dx = r \sec^2 \theta \, d\theta \tag{v}$$

Using (iv) and (v) in (iii), we get

$$E = \frac{\lambda}{2\pi\epsilon_0 r} \int_{\theta=0}^{\theta=\pi/2} \cos \theta \, d\theta$$

Notice that we have changed the limits of integration on x to θ. When $x = 0$, $\theta = 0$ and when $x = \infty$, $\theta = \pi/2$ as is obvious from (iv). The above equation on integration gives

$$E = \frac{\lambda}{2\pi\epsilon_0 r} \cdot \sin \theta \Big|_0^{\pi/2} = \frac{\lambda}{2\pi\epsilon_0 r}$$

which is Eq. (2.42). This example clearly illustrates the usefulness of Gauss's theorem over Coulomb's law. Notice that the calculation of E based on Coulomb's law is far more complicated than that based on Gauss's law.

EXAMPLE 2.13 Two parallel plane sheets 1 and 2 carry uniform charge densities σ_1 and σ_2 respectively. Find the electric field in each of the three regions marked I (to the left of both), II (between them) and III (to the right of both) as shown in Fig. 2.23. Discuss the special case when $\sigma_1 = -\sigma_2 = \sigma$.

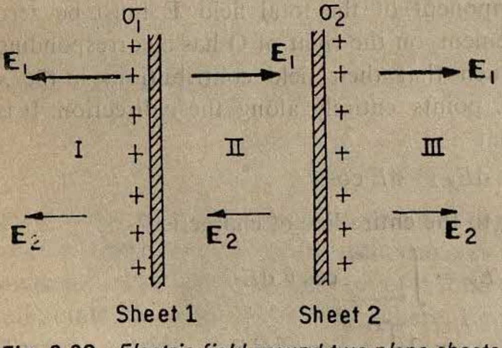


Fig. 2.23 *Electric field around two plane sheets* $(\sigma_1 > \sigma_2)$

Solution: We can find the electric field in each region by using Eq. (2.43) and the principle of superposition. The left sheet 1 produces a field of magnitude $E_1 = \sigma_1/2\epsilon_0$ which points away from it, to the left in region I and to the right in regions II and III. The right sheet 2 produces a field of magnitude $E_2 = \sigma_2/2\epsilon_0$ which points away from it; to the left in regions II and I and to the right in region III. If $\sigma_1 > \sigma_2 > 0$ the field in each region from the principle of superposition will be as follows:

Region I: $E = -E_1 - E_2 = -\dfrac{1}{2\epsilon_0}(\sigma_1 + \sigma_2)$

Region II: $E = E_1 - E_2 = \dfrac{1}{2\epsilon_0}(\sigma_1 - \sigma_2)$

Region III: $E = E_1 + E_2 = \dfrac{1}{2\epsilon_0}(\sigma_1 + \sigma_2)$

A negative field indicates that it points from right to left.

In the special case when $\sigma_1 = -\sigma_2 = \sigma$, i.e. the two sheets carry equal and opposite charge densities, then $\sigma_1 + \sigma_2 = 0$ and $\sigma_1 - \sigma_2 = 2\sigma$. Thus the fields in the regions I and III will be zero and that in region II will be σ/ϵ_0, pointing to the right between the sheets. Thus in this case, the field exists only in the region between the sheets and is zero elsewhere. The field between the sheets is uniform and is independent of their separation. In practice, this is the way to produce a uniform field in a limited region of space; the only other requirement is that the sheets should be much larger in linear dimension than in this region.

2.7 CONDUCTORS AND INSULATORS

So far we have studied electrostatic fields due to various kinds of charge distributions in which the charges were placed at definite locations and were *held there at rest* by some means. From these distributions we have learnt some basic properties of electrostatic fields. We will now discuss the electrical properties of materials in bulk.

Materials are broadly classified into *conductors* and *insulators*. In Chapter 1, we have briefly described the distinction between them in terms of free and bound electrons. In an *insulator*, such as glass or rubber, each electron is *attached* (or *bound*) to a particular atom and hence is not free to move in the body of the material. Insulators are also called *dielectrics*. On the other hand, in a metallic *conductor*, one or more electrons per atom are free to roam about through the body of the material. They cannot leave the conductor under normal circumstances. In liquid conductors, such as salt solution or copper sulphate solution, it is the ions that move. A *perfect* conductor would be a material which contains an *unlimited* supply of completely free charges. An insulator does not have such freely movable charges as each electron stays near its 'parent' nucleus or within its atom or molecule and cannot move far away from it. In terms of this distinction between conductors and insulators we will now discuss some of their electrical properties.

Electrical Properties of Dielectrics (or Insulators)

While electrons in dielectrics stay attached to specific molecules (or atoms), they can move within the molecule. The cumulative effect of such microscopic movements accounts for the characteristic behaviour of dielectric materials. Some examples are glass, wax, water, oil, air, wood, rubber, stone, plastics, etc.

What happens to a neutral atom when it is placed in an electric field? Our first guess is that nothing happens to it; since the atom is not charged, the field has no effect on it. But that is a wrong guess. Although the atom as a whole is electrically neutral, there is a positively charged nucleus and a negatively charged electron cloud surrounding it. These two regions of charge have their own 'centres of charge' within each molecule. The positive centre of charge is the centre of mass of the protons in the nuclei and the negative centre of charge is the centre of mass of the electrons in the molecule.

Non-polar and Polar Dielectrics In some dielectrics, in the absence of an external electric field, the two centres of charge in each molecule coincide. Each molecule, therefore, has a zero dipole moment in its normal condition. Such molecules are called *non-polar molecules* and the dielectrics made of such molecules are called non-polar dielectrics.

In other dielectrics, even in the absence of an external electric field, the two centres of charge in each molecule do not coincide because of the asymmetric shape of the molecule. Such molecules are called *polar molecules*; they have built-in *permanent dipole moments.* Some examples of polar molecules are water, ammonia, carbon monoxide, hydrogen chloride and methanol. The water molecule has a dipole moment because it is bent in the middle, the O–H axes making an angle about 105° with one another. The dipole moment of water is unusually large; about 6.1×10^{-30} Cm, which accounts for its effectiveness as a good solvent. Substances consisting of polar molecules are called polar dielectrics. We will not discuss polar dielectrics in any detail.

Polarization: When a dielectric is placed in an external electric field **E**, a small *induced dipole moment* is created in each molecule of even nonpolar dielectrics. This is because the positive charge (protons) in each atom is pushed in the direction of the field and the negative charge (electrons) is pushed in the opposite direction. In principle, if the field is large enough, it can pull the atom apart completely, i.e. ionize it. With moderate fields, however, an equilibrium is soon established, for if the centre of the negative charge does not coincide with that of the positive charge, these opposite charges attract each other, and this holds the atom together. The two opposing forces—(i) **E** pulling the electrons and the nucleus apart and (ii) their mutual attraction pushing them towards each other—quickly reach a balance (equilibrium), leaving the atom *polarized*, with the positive charge shifted slightly one way and the negative charge the other way. The atom now has a tiny dipole moment which points in the direction of **E**. The effects of such induced dipole moments or polarization will be described in Sec. 2.9.

Electrical Properties of Conductors

The general features of the behaviour of conductors in an electrostatic field are very different from those of dielectrics. The reason is that conductors have free charges but dielectrics do not. The word electrostatic means the charges are *static* implying that there is no movement of charges (i.e. there is no current) in the interior or on the surface of the conductor. Under electrostatic conditions, conductors have the following electrical properties.

1. *There is no net electric field in the interior of a conductor.* If there were any field, the free charges would move and the conditions would not be electrostatic any more. Is this a satisfactory explanation? Obviously not. All that it proves is that we cannot have electrostatics when conductors are present. We must, therefore, examine what happens when a conductor is placed in an external electric field $\mathbf{E}_{ext}$. Initially the field will push the positive charges to the right and the negative charges to the left (Fig. 2.24). (In metallic conductors it is only the negative charges—the electrons—that do the moving, the massive protons remain practically at rest. But when the electrons depart from the right side, this side is left with a net positive charge—the stationary protons. So it does not really matter which charges actually move; the effect is the same. In electrolytic solutions, on

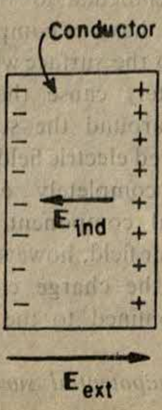


Fig. 2.24

other hand, it is the ions that do the moving and both positive and negative ions move.) When the charges reach the end of the material, they pile up; the positive charge on the right side and the negative charge on the left. Now these so called 'induced' charges produce an electric field $\mathbf{E}_{ind}$ of their own, which is in the direction *opposite* to that of the external electric field $\mathbf{E}_{ext}$. This is the crucial point. Thus the field of the induced charges tends to cancel the original field. Charges will continue to move until this cancellation is complete, i.e. until $\mathbf{E}_{ind}$ and $\mathbf{E}_{ext}$ became equal in magnitude. When this happens, electrostatic conditions are restored and the net electric field *inside* the conductor is zero. It may be mentioned that the whole process is practically instantaneous. In copper, for example, this process of rearrangement of charges takes as short a time as 10^{-17} s. It may also be borne in mind that *outside* the conductor the field is *not* zero, for here $\mathbf{E}_{ext}$ and $\mathbf{E}_{ind}$ do not cancel each other.

2. *There is no net charge in the interior of a conductor.* Since the electric field cannot exist in the interior of a conductor (i.e. $\mathbf{E} = 0$) the surface integral $\int \mathbf{E} \cdot d\mathbf{S}$ over the closed surface bounding *any* volume *inside* the conductor is zero. It then follows from Gauss's theorem (2.40) that $q = 0$, i.e. there can be no charge anywhere in the interior of the conductor.

3. *Charge resides on the surface of a conductor.* If a conductor is charged, i.e. some excess charge is added from outside, that extra charge can only reside on the surface of the conductor.

4. *Just outside a conductor, the electric field is perpendicular to the surface.* If the electric field were not perpendicular to the surface, it would have a component tangential to the surface which would immediately cause the flow of charges around the surface until the induced electric field due to the charges completely cancels the tangential component. The perpendicular field, however, remains because the charge cannot flow perpendicular to the surface since it is confined to the conductor (Fig. 2.25).

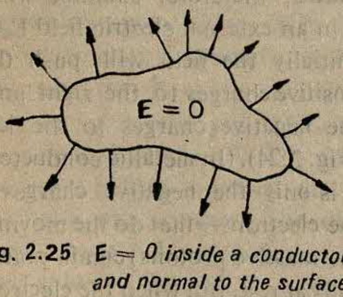


Fig. 2.25 $\mathbf{E} = 0$ *inside a conductor and normal to the surface outside*

5. *The surface of a conductor is an equipotential surface.* Since $\mathbf{E} = 0$, the line integral $\int_A^B \mathbf{E} \cdot d\mathbf{l} = 0$ between any two points A and

B within or at the surface of a conductor. Hence the potential difference between any two points is zero. Thus the potential is constant within and at the surface of a conductor (though different on different conductors). In particular, the surface of a conductor is an equipotential surface.

6. *The surface charge density is high where the radius of curvature of the surface of the conductor is small.* The surface region of a conductor is usually some two atoms thick, i.e. about 2×10^{-10} m (2Å). The surface charge density σ need not be constant. If some extra charge is given to a conductor (which was previously neutral), this charge spreads on the surface. The local charge density σ will, in general, not be the same everywhere on the surface; it will be positive at some places and negative at others. The charge density tends to be relatively high on sharp points and relatively low on plane regions on a conducting surface. The electric field **E** at points in space immediately above a charged surface is proportional to the charge density σ there so that **E** assumes very high values near sharp points. Glow discharges from sharp points during thunderstorms are a familiar example. The lightning conductor acts in this way to neutralize charged clouds and thus prevent lightning strokes.

Electrostatic Shielding The property of zero net field within a conductor is of great practical importance where the effects of external fields are undesirable. Thus any region of space can be isolated from external fields by surrounding it by a conducting surface. If there is no charge inside such a metal shield, the field inside the volume will be zero irrespective of any field outside the shield. To prevent electric interaction, the neighbouring circuits in electronic devices are shielded by placing metal shields around certain components or circuits.

EXAMPLE 2.14 An uncharged conductor has a cavity inside it. Somewhere in the cavity is a charge $+q$. What can you say about the electric field (a) inside the cavity, (b) within the conductor, and (c) outside the conductor?

Solution: (a) The cavity is electrically isolated from the conductor surrounding it. Since it has a charge, the electric field inside the cavity is nonzero.

(b) The charge $+q$ induces an opposite charge $-q$ on the wall of the cavity which distributes itself in such a way that its field cancels that of $+q$ for all points in the conductor (in fact, for all points exterior to the cavity). Hence there is no net electric field within the conductor (see Fig. 2.26).

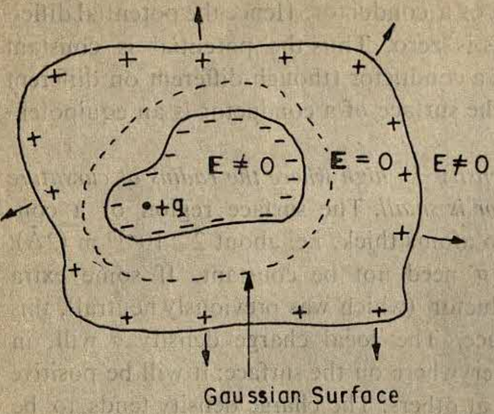


Fig. 2.26

(c) Since the conductor carries no net charge, this leaves charge $+q$ to distribute itself over the surface of the conductor. For all points outside the conductor, the only field that survives is the field of the leftover charge $+q$ on the outer surface. This charge produces a non-zero field in the region surrounding the conductor. To show that a charge $-q$ is induced on the inner surface of the cavity, surround the cavity with a Gaussian surface as shown in the figure. Since every point on this surface is inside the conductor, the surface integral of **E** over this surface vanishes and hence, by Gauss's theorem, the charge enclosed by this surface is zero, i.e. $q_{\text{enc}} = 0$. But $q_{\text{enc}} = q + q_{\text{ind}}$ where q_{ind} is the charge induced on the inner surface. So $q_{\text{ind}} = -q$.

Note: Coulomb's law and Gauss's theorem are complementary to each other; one cannot be true if the other does not hold; Coulomb's law being the more fundamental of the two. Electrostatic shielding, a consequence of Gauss's theorem, has, therefore, been used to test the accuracy of the exact inverse square dependence on distance in Coulomb's law $(1/r^2)$. If experiments (whose accuracy and sensitivity are finite) show that the field inside a conductor is zero, then the $1/r^2$ dependence in Coulomb's law can be ascertained to the accuracy of the experiment. The most delicate experiments have shown that, for r in the range of a few centimetres right down to atomic distances of r of up to 1 Å ($=10^{-10}$ m) the departure from $1/r^2$ dependence is less than 1 in 10^9. Thus Coulomb's law has been extremely well tested. But for distances of the order of 10^{-14} cm (nuclear dimensions) there are significant departures which are due to quantum mechanical effects.

2.8 CAPACITORS AND CAPACITANCE

We have seen how we can compute the electric field due to an arrangement of charges in space. We have also discussed the work per unit charge to move a test charge from one place to another, i.e. the potential difference between two points in the given field. We will now be interested in the charge distribution itself and in the work necessary to establish it. In particular, we are concerned with arrangements of one or more conducting bodies on which charges can be placed. Such equipotential surfaces are of great practical importance and have the virtue that the relationship between the amount of charge on each body and its potential is remarkably simple. The arrangement of most practical importance is that involving two conducting bodies separated by a certain distance. Such as arrangement is a new circuit element called the *capacitor* (or condenser).

General Concept of a Capacitor

Any isolated system of two conducting bodies, of any shape and size, separated by a distance in space, is called a capacitor.

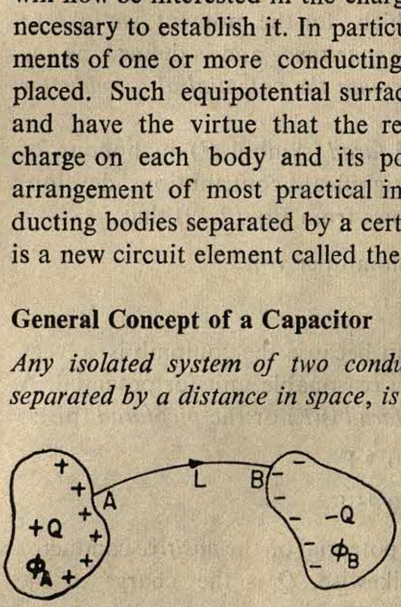


Fig. 2.27 *General concept of a capacitor*

Suppose we have two conductors A and B and we put a charge $+Q$ on A and $-Q$ on B. These charges will distribute themselves on the surfaces of A and B so that each surface is an equipotential surface. Let ϕ_A and ϕ_B be their respective constant potentials. The charges will produce a definite electric field **E** in the space around the conductors. The potential difference V between them can be calculated as the line integral of **E** from *any* point on the surface of A to *any* point on the surface of B along *any* path in space (one such path is shown in the figure).
Thus

$$V = \phi_A - \phi_B = \int_A^B \mathbf{E}\cdot d\mathbf{l}$$

It follows from Eq. (2.46) that E is proportional to Q, i.e. if Q is doubled, **E** is also doubled. This implies that V is proportional to Q, or

$$Q \propto V$$

or $$Q = CV \qquad (2.49)$$

where C is a constant of proportionality and its value depends upon the size, shape, the separation of the conductors and the nature of the dielectric medium between them which, for the moment, we take to be a vacuum. This constant is called *capacitance* and is defined as *the amount of charge in* C *necessary to increase the potential of a conductor by 1 V.*

The SI unit of capacitance is called *farad* (symbol F) in honour of Michael Faraday who, among other significant contributions, developed the concept of capacitance. If follows from Eq. (2.49) that

$$1 \text{ farad} = 1 \text{ coulomb per volt}$$

or $$1F = 1 \text{ CV}^{-1}$$

The farad is an enormously large unit because the coulomb is a very big unit of charge; the volt being a reasonable unit for potential. A more practical unit of C is the *microfarad* (μF) or the *picofarad* (pF):

$$1 \ \mu\text{F} = 10^{-6} \text{ F}$$

$$1 \text{ pF} = 10^{-12} \text{ F}$$

Notice that, by definition, V is the potential of the *positive* conductor minus that of the negative one. Likewise, Q is the charge of the *positive* conductor so that the capacitance is an intrinsically positive quantity. By the way, a *single* conductor may also have a capacitance. In this case the 'second conductor' is at infinity, i.e. V is the potential of the conductor when it carries a charge Q, using infinity as the reference point where the potential is zero. In practice, however, the earth is taken to be at zero potential.

Calculating Capacitance

We will now describe a few examples of capacitors in which the two conductors have simple geometrical shapes and find the capacitance in each case.

(*a*) *Parallel-plate Capacitor* The simplest example is the parallel-plate capacitor which consists of two parallel conducting plates, each of area A and separated by a distance d (Fig. 2.28).

We give a charge of $+Q$ to the upper plate and $-Q$ to the lower plate. (This can be done by connecting the upper plate to the positive

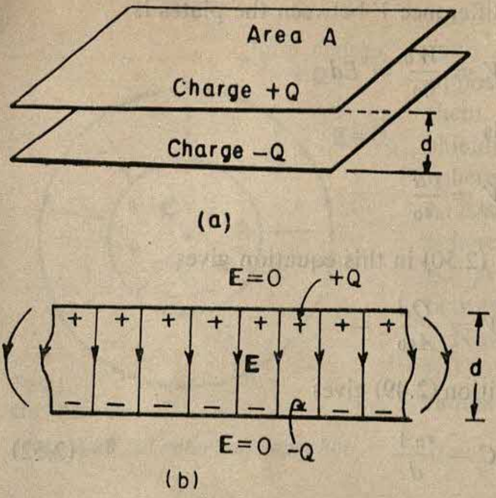


Fig. 2.28 (a) *A parallel-plate capacitor (b) Cross-section showing lines of force*

terminal of a battery and the lower plate to the negative terminal.) The charge will be distributed uniformly on the surface of each plate and a uniform surface charge density $+\sigma$ is established on the upper plate and $-\sigma$ on the lower plate. If the separation d is small compared

$$\sigma = \frac{Q}{A} \tag{2.50}$$

with the plate dimensions, the electric field **E** between the plates will be essentially uniform with some deviations from uniformity at the edges of the plates. Thus the lines of force will be parallel and evenly spaced as shown, with some distortion at the edges. For $d \ll A$, the edge-effects may be ignored.

From Gauss's theorem it follows that $\mathbf{E} = 0$ in the region outside the plates and the magnitude of the field in the region between the plates is (see Example 2.13 on page 90).

$$E = \frac{\sigma}{\epsilon_0} \tag{2.51}$$

Now the work W_0 required to carry a test charge q_0 from one plate to the other is equal to the product of the force q_0E and the distance d, i.e.

$$W_0 = q_0Ed$$

Therefore, the potential difference V between the plates is

$$V = \frac{W_0}{q_0} = Ed$$

which from Eq. (2.51) yields

$$V = \frac{\sigma d}{\epsilon_0}$$

Substituting for σ from Eq. (2.50) in this equation gives

$$V = \frac{Qd}{A\epsilon_0}$$

which with the use of definition (2.49) gives

$$C = \frac{\epsilon_0 A}{d} \tag{2.52}$$

Notice that the capacitance is determined purely by the geometry of the capacitor.

Note: Equation (2.52) suggests a method for measuring ϵ_0, the permittivity of free space. A capacitor of accurately known plate area A and plate spacing d is made and its capacity C is determined by measuring charge Q and potential difference $V (C = Q/V)$. Knowing C, A and d, the value of ϵ_0 is calculated from relation (2.52). Incidentally this relation gives another the SI unit of ϵ_0 which is farad/metre (Fm^{-1}).

EXAMPLE 2.15 The parallel plates of an air-filled capacitor are 5.0 mm apart. What must be the plate area if the capacitance is to be 1 F?

Solution: $C = 1\text{ F},\ d = 5.0\text{ mm} = 5.0 \times 10^{-3}\text{ m}$

$$\epsilon_0 = 8.85 \times 10^{-12}\ \text{Fm}^{-1} \text{ (for air)}$$

$$\text{Plate area A} = \frac{Cd}{\epsilon_0}$$

$$= \frac{1 \times 5.0 \times 10^{-3}}{8.85 \times 10^{-12}}$$

$$= 5.6 \times 10^{8}\ \text{m}^2$$

If we have a square plate, the side of the square is

$$\sqrt{5.6 \times 10^{8}} = 2.4 \times 10^{4}\ \text{m} = 24\ \text{km}$$

Thus a farad is indeed a large unit.

(*b*) *Spherical Capacitor* It consists of two concentric spherical conducting shells of inner and outer radii r_a and r_b with equal and opposite charges $\pm Q$ spread over them. Because of electrostatic shielding the field within the inner sphere is zero. i.e. $\mathbf{E} = 0$ for $r < r_a$. Also the field outside the outer sphere is zero, i.e. $\mathbf{E} = 0$ for $r > r_b$. A radial field exists only between the two spheres.

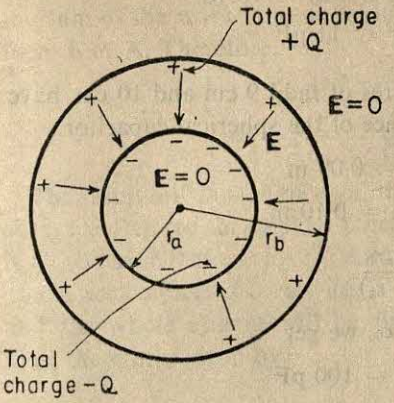


Fig. 2.29 *A spherical capacitor*

Now we know that the potential of a conducting sphere of radius r carrying a charge Q is

$$\phi = \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{r}$$

Hence the potential difference between the two spheres is

$$V = \frac{Q}{4\pi\epsilon_0}\left(\frac{1}{r_a} - \frac{1}{r_b}\right)$$

The capacitance of the spherical capacitor is given by

$$C = \frac{Q}{V} = 4\pi\epsilon_0 \frac{r_a r_b}{(r_b - r_a)} \qquad (2.53)$$

(*c*) *An Isolated Conducting Sphere* Suppose a conducting sphere of radius r is given a charge Q. Then the potential on the surface of the sphere is

$$V = \frac{Q}{4\pi\epsilon_0 r}$$

$$\therefore \qquad \frac{Q}{V} = 4\pi\epsilon_0 r$$

or

$$C = 4\pi\epsilon_0 r \qquad (2.54)$$

For a calculation of the 'capacitance of a cylindrical capacitor', refer to Supplement S3 at the end of this chapter.

EXAMPLE 2.16 Find the capacitance of a conducting sphere of radius 10 cm situated in air.

Solution: $r = 10 \text{ cm} = 0.10 \text{ m}$

$$\epsilon_0 = 8.85 \times 10^{-12} \text{ Fm}^{-1}$$

$$C = 4\pi\epsilon_0 r$$
$$= 4 \times 3.14 \times 8.85 \times 10^{-12} \times 0.10$$
$$= 11 \times 10^{-12}\ \text{F} = 11\ \text{pF}$$

EXAMPLE 2.17 Two concentric spheres of radii 9 cm and 10 cm have air between them. Find the capacitance of the spherical capacitor.

Solution: $r_a = 9\ \text{cm} = 0.09\ \text{m}$

$$r_b = 10\ \text{cm} = 0.10\ \text{m}$$

$$C = \frac{4\pi\epsilon_0 r_a r_b}{(r_b - r_a)}$$

Substituting the values of r_a, r_b and ϵ_0, we get

$$C = 10^{-10}\ \ \text{F} = 100\ \text{pF}$$

Note that the inclusion of a second sphere increases the capacitance. For an isolated sphere of radius 10 cm, the capacitance was only 11 pF (Ex. 2.16).

Combinations of Capacitors

In many circuits, especially in radio and television sets, capacitors often appear in combinations whose resultant capacitance must be known. The two most common combinations involve connections of capacitors in *series* and in *parallel.*

(*a*) *Capacitors in Series* Figure 2.30 shows three capacitors of capacitances C_1, C_2 and C_3 connected in series. The right-hand plate of one is connected to the left-hand plate of the other and so on.

When a voltage V is applied across the free ends A and G of the combination (this can be done by connecting a cell across these two ends), a charge, say Q, is transferred from plate G to plate A; a charge $-Q$ being left on G (because it was neutral earlier). This charge $-Q$ induces a charge $+Q$ on plate F. Similarly charges appear on all the other capacitor plates, as shown in the figure. The potential differences across the individual capacitors are, therefore, given by

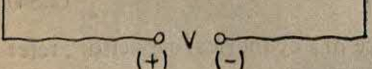


Fig. 2.30 *Capacitors in series*

$$V_1 = \frac{Q}{C_1} \qquad V_2 = \frac{Q}{C_2} \qquad V_3 = \frac{Q}{C_3} \tag{2.55}$$

The sum of these must be equal to the applied potential difference V because the work done in taking a test charge from G to A is the sum of the work done in taking it from G to F, from E to D and from B to A. Therefore,

$$V = V_1 + V_2 + V_3$$
$$= Q\left(\frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3}\right) \qquad (2.56)$$

The resultant capacitance of the combination is the ratio of the charge stored to the applied potential difference V. The charge stored is Q, because if the source (i.e. the cell) is removed and the plates G and A are connected by a wire, a charge Q will pass through the wire and the whole system will be discharged.. The resultant capacitance C is, therefore, given by

$$C = \frac{Q}{V} \quad \text{or} \quad \frac{1}{C} = \frac{V}{Q}$$

which, with use of Eq. (2.56), becomes

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \ldots \qquad (2.57)$$

Thus *the reciprocal of the resultant capacitance of capacitors in series is equal to the sum of the reciprocals of the individual capacitances.* The resultant capacitance C is less than the smallest among $C_1, C_2, C_3, \ldots$.

In general, the resultant capacitance of N capacitors in series is

$$\frac{1}{C} = \sum_{n=1}^{N} \frac{1}{C_n}$$

(*b*) *Capacitors in Parallel* Figure 2.31 shows three capacitors of capacitances C_1, C_2, and C_3 connected in parallel. All their left-hand plates are connected together, as also are all their right-hand plates. If a cell of voltage V is connected across them as shown, they all will have a common potential difference V but different charges given by

$$Q_1 = C_1V \qquad Q_2 = C_2V$$
$$Q_3 = C_3V$$

The total charge stored in the system is

$$Q = Q_1 + Q_2 + Q_3$$
$$= V(C_1 + C_2 + C_3)$$

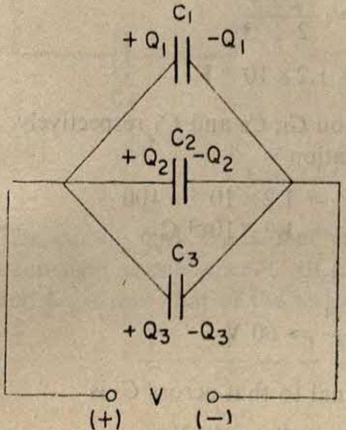


Fig. 2.31 *Capacitors in parallel*

The resultant capacitance of the combination is

$$C = \frac{Q}{V} = C_1 + C_2 + C_3 \ldots \qquad (2.58)$$

or

$$C = \sum_{n=1}^{N} C_n$$

Thus *the resultant capacitance of capacitors in parallel is equal to the sum of the individual capacitances.* The resultant capacitance is greater than the greatest among C_1, C_2, C_3, . . .

EXAMPLE 2.18 Three capacitors of capacitances $C_1 = 2\mu\text{ F}, C_2 = 2\mu\text{ F}$ and $C_3 = 1\mu\text{ F}$ are connected to a 100 V supply as shown in Fig. 2.32. Find (a) the equivalent capacitance of the combination and (b) the charges on each capacitor.

Solution: (a) The capacitance of the parallel combination of C_2 and C_3 is

$$C' = C_2 + C_3 = 2\mu\text{ F} + 1\mu\text{ F} = 3\mu\text{ F}$$

This parallel combination is connected in series to C_1. Hence the equivalent capacitance C of the entire combination is given by

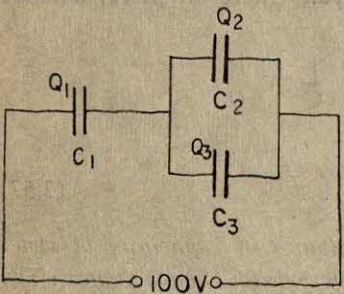


Fig. 2.32

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C'}$$

or

$$C = \frac{C_1 C'}{C_1 + C'} = \frac{2 \times 3}{2 + 3}$$

$$= 1.2\mu\text{ F} = 1.2 \times 10^{-6}\text{ F}$$

(b) Let Q_1, Q_2 and Q_3 be the charges on C_1, C_2 and C_3 respectively. The total charge Q stored in the combination is

$$Q = Q_1 = Q_2 + Q_3 = CV = 1.2 \times 10^{-6} \times 100$$
$$= 1.2 \times 10^{-4}\text{ C}$$

Therefore, potential difference across C_1 is

$$V_1 = \frac{Q_1}{C_1} = \frac{1.2 \times 10^{-4}}{2 \times 10^{-6}} = 60\text{ V}$$

So, the potential difference across C_2 equal to that across C_3 is

$$V_2 = V_3 = V - V_1 = 100 - 60 = 40\text{ V}$$

Hence $Q_2 = C_2V_2 = 2\times10^{-6}\times40 = 8\times10^{-5}$ C

and $Q_3 = C_3V_3 = 1\times10^{-6}\times40 = 4\times10^{-5}$ C

EXAMPLE 2.19 (a) Find the equivalent capacitance of the network shown in Fig. 2.33. (b) For a 300 V supply, determine the voltage across each capacitor.

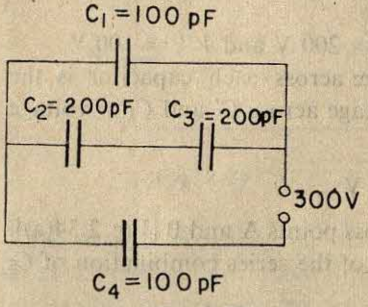


Fig. 2.33

Solution: (a) The series combination of C_2 and C_3 is equivalent to a capacitance C' given by

$$\frac{1}{C'} = \frac{1}{C_2} + \frac{1}{C_3}$$

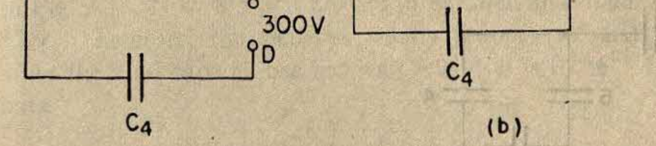
or $$C' = \frac{C_2C_3}{C_2 + C_3}$$

$$= \frac{200\times200}{200+200}$$

$$= 100 \text{ pF}$$

Therefore the circuit reduces to the one shown in Fig. 2.34(a). The equivalent capacitance between points A and B is

$$C'' = C_1 + C' = 100 + 100 = 200 \text{ pF}$$



(a) (b)

Fig. 2.34

The circuit may be further simplified to that in Fig. 2.34(b). The equivalent capacitance C of the entire network, i.e. between points A and D, is now that of the series combination of C'' and C_4. Thus

$$\frac{1}{C} = \frac{1}{C''} + \frac{1}{C_4} = \frac{1}{200} + \frac{1}{100} = \frac{3}{200}$$

or $$C = \frac{200}{3} \text{ pF}$$

(b) In the series combinations, the voltage divides itself among the capacitors in the inverse proportion of their capacitances (see Fig. 2.55). Hence the voltages across C'' and C_4 in Fig. 2.34(b) are in the ratio

$$\frac{V''}{V_4} = \frac{C_3}{C''} = \frac{100}{200} = \frac{1}{2} \quad \text{or} \quad V_4 = 2V''$$

But $V'' + V_4 = 300$ V. Therefore $V_4 = 200$ V and $V'' = 100$ V.

In a parallel combination, the voltage across each capacitor is the same. Referring to Fig. 2.34(a) the voltage across C' and C_1 = voltage across C''. Hence.

$$V_1 = 100 \text{ V}$$

and $V' = 100$ V; this is the voltage across points A and B [Fig. 2.34(a)]. Now C' is the combined capacitance of the series combination of C_2 and C_3. Hence

$$\frac{V_2}{V_3} = \frac{C_3}{C_2} = \frac{200}{200} = 1$$

But $\qquad V_2 + V_3 = V' = 100$ V

Hence $\qquad V_2 = 50$ V and $V_3 = 50$ V

Thus $\quad V_1 = 100$ V, $V_2 = 50$ V, $V_3 = 50$ V and $V_4 = 200$ V.

EXAMPLE 2.20 Figure 2.35 shows a network of capacitors where the numbers indicate capacitances in microfarads. Find the value of capacitance C if the equivalent capacitance between points A and B is to be 1 μF.

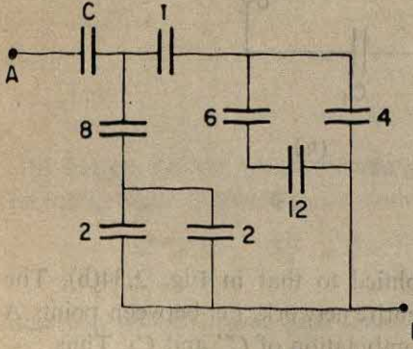


Fig. 2.35

Solution: The series combination of 6 and 12 is equivalent to 4 and the parallel combination of 2 and 2 is also equivalent to 4. Therefore the network can be simplified as shown in Fig. 2.36.

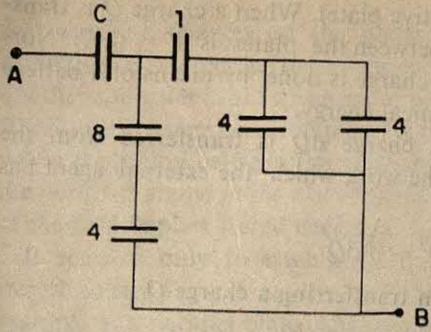


Fig. 2.36

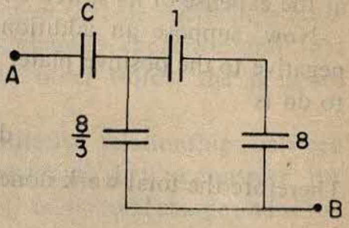


Fig. 2.37

The parallel combination of 4 and 4 is equivalent to 8 and the series combination of 8 and 4 is equivalent to 8/3. Thus the combination in Fig. 2.36 reduces to that in Fig. 2.37.

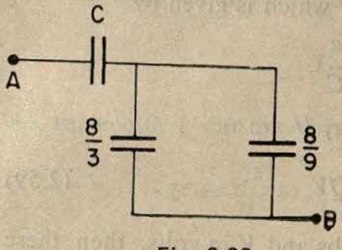


Fig. 2.38

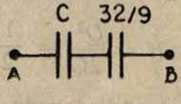


Fig. 2.39

The series combination of 1 and 8 in Fig. 2.37 yields 8/9 as shown in Fig. 2.38. Now 8/3 and 8/9 are in parallel and their equivalent is 32/9. Therefore, the network finally reduces to that in Fig. 2.39. Since the total capacitance between A and B is to be 1 (i.e. 1μF), we have

$$1 = \frac{1}{C} + \frac{9}{32}$$

which gives

$$C = \frac{32}{23}\,\mu\text{F} = 1.39\ \mu\text{F}$$

Energy Storage in a Capacitor

A charged capacitor is a store of electrical energy. We can visualize the 'charging' of a capacitor by imagining that an external agent pulls electrons from the 'positive' plate and transfers them to the 'negative' plate (or equivalently, the external agent transfers positive charge from the negative to the positive plate). When a charge Q is transferred, the potential difference between the plates is $V = Q/C$. Normally the work of transferring charge is done by means of a battery at the expense of its stored chemical energy.

Now suppose an additional charge $\mathrm{d}Q$ is transferred from the negative to the positive plate, the work which the external agent has to do is

$$\mathrm{d}W = V\mathrm{d}Q$$

Therefore the total work done in transferring a charge Q is

$$W = \int_0^Q \mathrm{d}W = \int_0^Q V\mathrm{d}Q$$

But $V = Q/C$. Therefore,

$$W = \int_0^Q \frac{Q}{C}\,\mathrm{d}Q = \frac{1}{C}\int_0^Q Q\mathrm{d}Q = \frac{Q^2}{2C}$$

This work is stored as potential energy U which is given by

$$U(=W) = \frac{Q^2}{2C}$$

Since $Q = CV$, the expressions for energy U are

$$U = \frac{Q^2}{2C} = \frac{1}{2}CV^2 = \frac{1}{2}QV \qquad (2.59)$$

If C is measured in farads, Q in coulombs and V in volts, then these formulae will give U in joules.

If many capacitors are combined (in series or in parallel) the total energy stored in *either* combination is equal to the sum of the energies stored in individual capacitors. This follows from expressions (2.57) and (2.58) along with Eq. (2.59) as we show below:

(a) *Series combination*: $Q =$ constant

$$\text{Total energy } U = \frac{Q^2}{2}\left(\frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \ldots\right) = \frac{Q^2}{2C}$$

$$= U_1 + U_2 + U_3 + \ldots$$

(b) *Parallel combination*: $V =$ constant

$$\text{Total energy } U = \frac{1}{2}V^2(C_1 + C_2 + C_3 + \ldots) = \frac{1}{2}CV^2$$

$$= U_1 + U_2 + U_3 + \ldots$$

Where is this energy stored? We have seen above that the work done in transferring charge from one plate to another (so as to have equal and opposite charges on the plates) against their mutual force of attraction is stored in the capacitor as potential energy. There is another equivalent (but much more physical) way of looking at the stored energy—one which emphasizes the fact that electric field is created in the region between the plates of a charged capacitor and that the *energy is stored in the electric field.* In other words, the presence of the field implies stored energy.

It remains only to establish a quantitative relationship between stored energy U and electric field strength E. Let us consider the example of a parallel plate capacitor. If Q is the total charge and A the plate area, the surface charge density is

$$\sigma = \frac{Q}{A} \qquad \text{or} \qquad Q = \sigma A$$

We know that the magnitude of the electric field (which, neglecting the edge effects, is uniform in the space between the plates) is given by

$$E = \frac{\sigma}{\epsilon_0} \tag{2.60}$$

So, the energy U can be written in terms of E as follows:

$$U = \frac{Q^2}{2C} = \frac{(\sigma A)^2}{2\epsilon_0 A/d} \qquad \left(\because C = \frac{\epsilon_0 A}{d}\right)$$

$$= \frac{1}{2}\frac{\sigma^2 Ad}{\epsilon_0} = \frac{1}{2}\epsilon_0 E^2 \cdot Ad$$

Now $Ad =$ volume V of the capacitor (between the two plates). Therefore the energy density (u) defined as total energy per unit volume is given by

$$u = \frac{U}{V} = \frac{1}{2}\epsilon_0 E^2 \tag{2.61}$$

Although this relation has been derived for the special case of a parallel plate capacitor, it is true in general. Thus we can quite generally state that *if an electric field* **E** *exists at any point in space then that point can be regarded as a site of stored energy whose amount per unit volume is* $\frac{1}{2}\epsilon_0 E^2$. Similarly, energy is also associated with a magnetic field.

Example 2.21 A capacitor of capacitance C_1 is charged by a potential difference V_0. The charging battery is then removed and the capacitor is connected to an uncharged capacitor of capacitance C_2 as shown in Fig. 2.40. (a) What is the common potential difference across the combination? (b) What is the stored energy of the system before and after the capacitors are connected? (c) Is there any loss in energy after connection? If so, what happens to the energy lost?

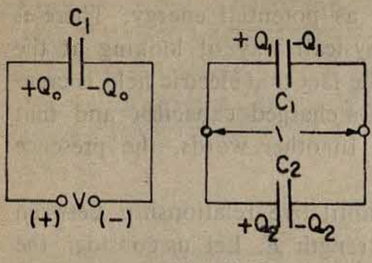


Fig. 2.40

Solution: (a) The original charge Q_0 is now shared by the two capacitors. Thus

$$Q_0 = Q_1 + Q_2$$

Using the relation $Q = CV$ we have

$$C_1V_0 = C_1V + C_2V$$

or

$$V = V_0 \frac{C_1}{C_1 + C_2} \tag{i}$$

(b) Since the second capacitor is uncharged, it has no energy. Therefore, the total energy of the system before connection is

$$U_0 = \text{energy of capacitor } C_1$$

$$= \frac{1}{2} C_1V_0^2 \tag{ii}$$

The total energy after connection is

$$U = \frac{1}{2} C_1V^2 + \frac{1}{2} C_2V^2$$

Using (i) we have

$$U = \frac{1}{2} (C_1 + C_2)\left(\frac{V_0C_1}{C_1 + C_2}\right)^2 \tag{iii}$$

(c) From (ii) and (iii) we find that

$$\frac{U}{U_0} = \frac{C_1}{C_1 + C_2}$$

which is less than unity. Hence there is a loss of energy after connection.

$$\text{Energy loss} = U_0 - U$$

$$= \frac{1}{2}\left(\frac{C_1C_2}{C_1 + C_2}\right) V_0^2$$

The 'missing' energy appears as heat in the connecting wires as the charges move through them when the capacitors are connected.

EXAMPLE 2.22 A 2μF capacitor charged to a potential difference of 100 V and a 5μF capacitor charged to a potential difference of 50V are connected together as shown in Fig. 2.41. Find the loss in energy due to connection.

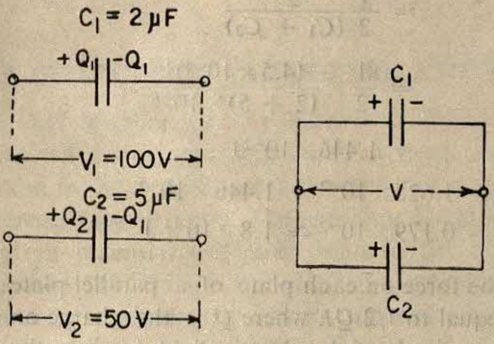


Fig. 2.41

Solution: Charge Q_1 on $C_1 = C_1V_1 = 2\times10^{-6}\times100$

$$= 2\times10^{-4}\text{C}$$

Charge Q_2 on $C_2 = C_2V_2 = 5\times10^{-6}\times50 = 2.5\times10^{-4}\text{C}$

$\therefore$ Total charge before connection $Q = Q_1 + Q_2$

$$= 4.5\times10^{-4}\text{C}$$

Total energy before connection $= \frac{1}{2}\,C_1V_1^2 + \frac{1}{2}\,C_2V_2^2$

$$= \frac{1}{2}\times2\times10^{-6}\times(100)^2$$

$$+\frac{1}{2}\times5\times10^{-6}\times(50)^2$$

$$= 1.625\times10^{-2}\text{ J}$$

When the capacitors are joined in parallel with plates of like charges connected together, some charge will flow from C_2 to C_1 (since $Q_2 > Q_1$) until the potential difference across each capacitor becomes equal to a common value V. Now, since the charge is conserved, the total charge on C_1 and C_2 after connection = total charge before connection = Q.

The common potential difference V after connection is given by

$$C_1V + C_2V = Q$$

or
$$V = \frac{Q}{C_1 + C_2}$$

$\therefore$ Total energy after connection $= \frac{1}{2}(C_1 + C_2)V^2$

$$= \frac{1}{2}\frac{Q^2}{(C_1 + C_2)}$$

$$= \frac{1}{2}\times\frac{(4.5\times10^{-4})^2}{(2+5)\times10^{-6}}$$

$$= 1.446\times10^{-2}\text{J}$$

$$\text{Loss of energy} = 1.625\times10^{-2} - 1.446\times10^{-2}$$

$$= 0.179\times10^{-2} \simeq 1.8\times10^{-3}\text{ J}$$

EXAMPLE 2.23 Show that the force on each plate of a parallel-plate capacitor has a magnitude equal to 1/2 QE where Q is the charge on the capacitor and E is the magnitude of the electric field between the plates.

Solution: To find the force between the charged capacitor plates we will use a method called the '*method of virtual displacement*'. We simply equate the work ΔW required to make a small change Δd in the plate separation d to the resulting change ΔU in the stored energy, i.e.

$$\Delta W = \Delta U \qquad \text{(i)}$$

If F is the magnitude of the force between the plates, then the work ΔW done to increase the plate separation is given by

$$\Delta W = F\Delta d \qquad \text{(ii)}$$

Now we know that the energy U of a parallel-plate capacitor of plate area A and capacitance C is

$$U = \frac{Q^2}{2C} = \frac{Q^2d}{2\epsilon_0A}$$

where Q is the charge on the capacitor plates. The increase ΔU in U due to an increase Δd in d is, therefore, given by

$$\Delta U = \frac{Q^2\Delta d}{2\epsilon_0A} \qquad \text{(iii)}$$

Equating (ii) and (iii) we get

$$F = \frac{Q^2}{2\epsilon_0A} \qquad \text{(iv)}$$

Now we know that the electric field E between the plates is given by [see Eq. (2.60)]

$$E = \frac{\sigma}{\epsilon_0} = \frac{Q}{\epsilon_0 A} \tag{v}$$

Using (v) in (iv) we get

$$F = \frac{1}{2} QE$$

2.9 DIELECTRICS AND POLARIZATION OF MATTER

In 1837 Michael Faraday discovered a remarkable effect. He showed that when an insulating material (such as glass or mica) was introduced in the space between the plates of a capacitor, its capacitance increased significantly. The capacitor could now hold more charge at a given potential difference than in the case when the space was a vacuum. The factor by which the capacitance is multiplied is called the *relative permittivity* ϵ_r (sometimes also called *dielectric constant*):

$$\epsilon_r = \frac{\text{capacitance of given capacitor with space between plates filled with dielectric}}{\text{capacitance of the same capacitor with plates in vacuum}}$$

If we take the case of a parallel plate capacitor, then

$$\epsilon_r = \frac{\epsilon A/d}{\epsilon_0 A/d} = \frac{\epsilon}{\epsilon_0} \tag{2.62}$$

where ϵ is the permittivity of the dielectric and ϵ_0 that of vacuum. Thus *the relative permittivity is the ratio of the permittivity of the substance to that of vacuum.* Note that ϵ_r is a pure number and has no dimensions, unlike ϵ and ϵ_0. Also ϵ_r is independent of the shape and size of the capacitor and its value varies widely for different substances (see Table 2.1).

Table 2.1 Relative permittivities (or dielectric constants) of some dielectrics

Substance	ϵ_r	Substance	ϵ_r
Vacuum	1.00000	Mica	6
Air	1.00054	Ebonite	2.8
Water	81	Porcelain	6.5
Paper	3.5	Bakelite	4.8
Glass	5–10	Ice	94
Amber	2.7	Paraffin wax	2

Note that, for air, the relative permittivity is so close to that of vacuum that we may neglect its effect for most purposes.

Polarization of Matter

We wish to understand how a dielectric increases the capacitance of a capacitor, i.e. what happens when a dielectric is placed in an electric field.

As mentioned in Sec. 2.7, there are two kinds of dielectrics—non-polar and polar. In their normal state non-polar atoms have no dipole moment. When such an atom is placed in a moderate electric field, the positive nucleus is pulled in one direction and the negative electrons in the opposite direction (Fig. 2.42). The atom is then said to be *polarized* (distorted). Thus two centres of charge are created and the atom acquires a dipole moment proportional to the applied field $\mathbf{E}_0$ (if the field is not too strong). This *induced dipole moment* $\mathbf{p}$ of a single atom may be written as

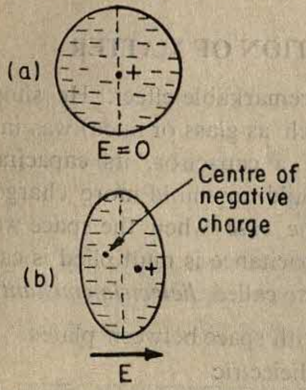


Fig. 2.42 *Polarization of an atom in an electric field* $\mathbf{E}_0$: *(a) a normal atom and (b) a polarized atom*

$$\mathbf{p} = \epsilon_0 \alpha E_0 \tag{2.63}$$

The proportionality constant α is called the *atomic polarizability* and its dimensions are

$$\frac{\text{dimensions of } p}{(\text{dimensions of } \epsilon_0)\times(\text{dimensions of } E_0)}$$

$$= \frac{\text{Cm}}{(\text{C}^2\text{N}^{-1}\text{m}^{-2})\times(\text{NC}^{-1})}$$

$$= \text{m}^3, \text{ the dimensions of volume}$$

For most atoms, α is of the order of 10^{-29} to 10^{-30} m^3, which is of the same order as atomic volumes. In such atoms, an external electric field of about of 10^6 Vm^{-1} will induce a dipole moment of order 10^{-34} Cm.

On the other hand, polar molecules, such as HCl, NH_3, CO_2 H_2O, etc. are permanently polarized even in the absence of an external field; they have permanent dipole moments of their own, typically in the range of about 10^{-30} Cm. For example, the water molecule consists of two hydrogen atoms and one oxygen atom making an angle of about 105° as shown in Fig. 2.43. The permanent dipole moment of a water molecule is about ten thousand times more than the induced moments in non-polar molecules. When a polar molecule is placed in an electric field, the creation of an additional dipole moment is not as important as the aligning of the existing moment along the direction of the field. We will not deal with polar molecules any further.

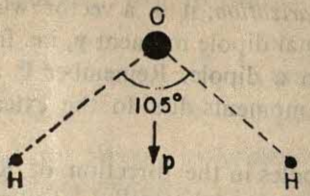


Fig. 2.43 *A water molecule*

Electric Polarization: Consider a rectangular slab of a (non-polar) dielectric. Let us imagine that all the atoms have been *uniformly* polarized, in a direction parallel to one of the edges of the slab by the application of a uniform external electric field $\mathbf{E}_0$, as shown in Fig. 2.44.

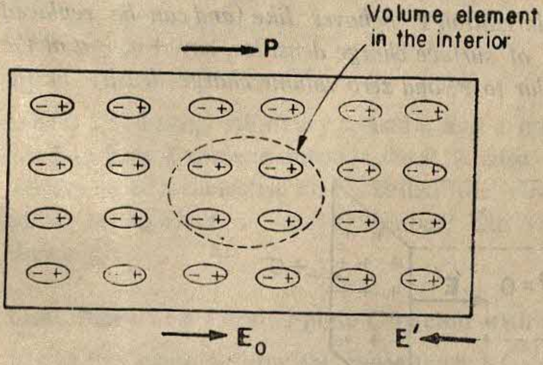


Fig. 2.44 *A uniformly polarized dielectric slab. Effective field* $\mathbf{E} = \mathbf{E}_0 - \mathbf{E}'$

Let us suppose that, within each atom, the positive and negative charges are separated by a vector displacement **x**. If q is the magnitude of positive or negative charge, the dipole moment of each atom is

$$\mathbf{p} = q\mathbf{x}$$

If N is the number density of atoms (i.e. the number of atoms per unit volume) the total microscopic dipole moment density is given by

$$\mathbf{P} = N\mathbf{p} = Nq\mathbf{x} \tag{2.64}$$

Since the unit of N is m^{-3}, the unit of **P** will be C m^{-2}. Quantity **P** is called *electric polarization* or simply *polarization*; it is a vector whose direction is the same as that of individual dipole moment **p**, i.e. from the negative to the positive charge, as in a dipole. Remember **P** is a cumulative effect of induced dipole moments due to the external field $\mathbf{E}_0$.

The overall effect of alignment of dipoles in the direction of $\mathbf{E}_0$ is to separate the centre of the positive charge of the entire slab slightly from the centre of the negative charge. The slab, as a whole, although remaining electrically neutral, is *polarized* uniformly, as Fig. 2.44 suggests. The net effect is pile-up of positive charge on the right face of the slab and of negative charge on the left face; but no net excess charge appears within any volume element (shown dotted) of the slab owing to the cancellation of positive and equal negative charges. Since the slab as a whole remains electrically neutral, the positive *induced surfaces charge* must be equal in magnitude to the negative induced surface charges on the two opposite sides of the slab. These two faces are perpendicular to electric polarization vector **P** (which is in the same direction as electric field vector $\mathbf{E}_0$). The net result is that, *for macroscopic purposes, a uniformly polarized rectangular dielectric slab with uniform electric polarization* **P** *behaves like (and can be replaced by) two surface layers of surface charge densities, say* $+\sigma$, $-\sigma$ *at the two surfaces perpendicular to* **P**, *and zero volume charge density in the interior* (see Fig. 2.45).

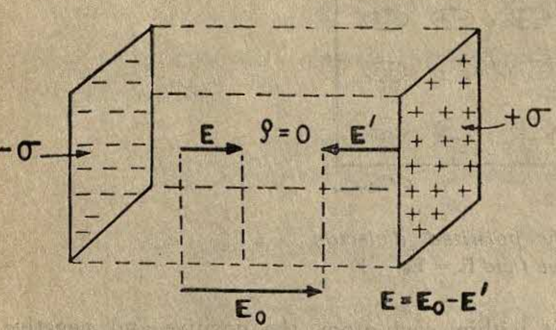


Fig. 2.45 *A polarized dielectric slab may be replaced by induced surface charges*

How is the surface charge density σ related to the magnitude P of electric polarization vector **P**? Consider a volume element of thickness x and cross-sectional area a, then the number of atomic dipoles in this volume element is Nax where N is the number per unit volume.

The induced charge in this volume is $Q = q\,Nax$ where q is charge on each dipole. Now, by definition,

$$\sigma = \frac{\text{charge}}{\text{area}} = \frac{Q}{a} = qNx$$

or $$\sigma = P \text{ [see Eq. (2.64)]}$$

Thus *the electric polarization may be defined as the induced surface charge per unit area.*

Notice from Fig. 2.45 that the induced surface charges will always appear in such a way that the electric field ($\mathbf{E}'$) set up by them opposes the external field $\mathbf{E}_0$, thus tending to weaken the original field within the dielectric. We will show below that it is this weakening of the field which is responsible for Faraday's discovery of increase in capacitance when a dielectric is introduced between the plates of a capacitor.

Electric Susceptibility We have seen that the effective electric field in a polarized dielectric is

$$\mathbf{E} = \mathbf{E}_0 - \mathbf{E}' \tag{2.65}$$

If $\mathbf{E}$ is not too large, the electric polarization $\mathbf{P}$ is proportional to $\mathbf{E}$, i.e.

$$\mathbf{P} \propto \mathbf{E}$$

or $$\mathbf{P} = \epsilon_0 \chi \mathbf{E} \tag{2.66}$$

where χ is a proportionality constant and a multiplicative factor ϵ_0 is used to keep χ dimensionless. Like P, χ also describes the electrical behaviour of a dielectric and is called the *electric susceptibility*. The value of χ (which is just a number) can vary widely for different dielectrics.

Capacitance of a Parallel-plate Capacitor with a Conducting Slab

Let us first examine how the capacitance of a parallel-plate capacitor changes when a conducting slab is introduced between its plates. Consider a capacitor of plate area A and plate separation d. If the space between the plates is a vacuum, its capacitance is given by

$$C_0 = \frac{\epsilon_0 A}{d} \tag{2.67}$$

What happens if a conducting slab of area A (equal to the plate area) and thickness $t < d$ is introduced between the plates?

Suppose the charges on the capacitor plates are $\pm Q$. When the conducting slab is introduced, as shown in Fig. 2.46, the free charges

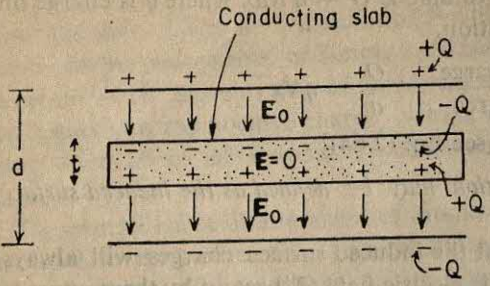


Fig. 2.46 *Introducing conducting slab in parallel-plate capacitor*

in the conducting slab flow so as to reduce the field to zero in the interior of the slab. Thus charges $\mp Q$ will appear on the two faces of the slab opposite the capacitor plates carrying charges $\pm Q$ respectively. The original uniform field $E_0 = Q/A\epsilon_0$ will now exist over a distance $(d - t)$ instead of d. Consequently, the potential difference between the capacitor plates *reduces* from the original (vacuum) value

$$V_0 = E_0 d = \frac{Qd}{A\epsilon_0}$$

to a new value

$$V = E_0(d - t) = \frac{Q}{A\epsilon_0}(d - t)$$

Hence the capacitance of the capacitor becomes

$$C = \frac{Q}{V} = \frac{\epsilon_0 A}{(d - t)} \tag{2.68}$$

Comparing this value with the vacuum value C_0 given by Eq. (2.67), we find that

$$C = \frac{C_0}{1 - (t/d)} \tag{2.69}$$

Thus $C > C_0$. The introduction of a conducting slab *increases* the capacitance. The reason for this increase is obvious. In a part $(=d-t)$ of the space between the plates, the field is reduced (to zero in the case of a conductor) which reduces the potential difference V. Since the total charge Q remains the same, it follows from $Q = CV$ that a decrease in V would result in an increase in capacitance C.

Capacitance of a Parallel-plate Capacitor with a Dielectric Slab

We will now examine how the capacitance of a parallel-plate capacitor changes when a dielectric slab is introduced between its plates. Consider a capacitor of plate area A and plate separation d. If the space between the plates is vacuum, its capacitance C_0 is given by Eq. (2.67). We insert a dielectric slab (instead of a conducting slab) of area A and thickness $t < d$ between the plates (Fig. 2.47).

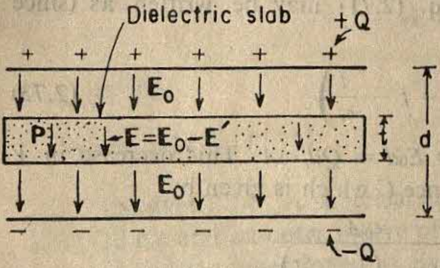


Fig. 2.47 *Introducing dielectric slab in parallel-plate capacitor*

Suppose initially the charges on the capacitor plates are $\pm Q$, which produce a uniform electric field $\mathbf{E}_0$ (neglecting edge effects) in the region between the plates. When a dielectric slab is introduced in this region, the field $\mathbf{E}_0$ polarizes the dielectric; the electric polarization vector P being in the same direction as $\mathbf{E}_0$ (see also Fig. 2.44). We have seen above that a polarized dielectric slab may be replaced by its induced surface charges for purposes of calculating the effect of the polarization on the field outside the slab (see Fig. 2.45). These induced charges produce a field $\mathbf{E}'$ of their own which is in the opposite direction to that of $\mathbf{E}_0$. Hence, *inside the dielectric* of thickness t, the effective field is

$$\mathbf{E} = \mathbf{E}_0 - \mathbf{E}'$$

Now the induced field $\mathbf{E}'$ is proportional to polarization vector $\mathbf{P}$, or $\mathbf{E}' \propto \mathbf{P}$ or $\mathbf{E}' = \mathbf{P}/\epsilon_0$ Thus

$$\mathbf{E} = \mathbf{E}_0 - \frac{\mathbf{P}}{\epsilon_0} \qquad (2.70)$$

But *outside the dielectric* the field remains equal to $\mathbf{E}_0$. Hence the potential difference V between the plates is given by

$$V = Et + E_0(d - t) \qquad (2.71)$$

In terms of electric susceptibility X defined by Eq. (2.66), Eq. (2.70) becomes

$$\mathbf{E} = \mathbf{E}_0 - X\mathbf{E}$$

or

$$(1 + X)\,\mathbf{E} = \mathbf{E}_0$$

or

$$(1 + X) = \frac{\mathbf{E}_0}{\mathbf{E}} = \epsilon_r \quad \text{(by definition)} \tag{2.72}$$

where ϵ_r is the relative permittivity (or dielectric constant) of the dielectric slab. In terms of ϵ_r, Eq. (2.71) may be written as (since $E_0 = Q/\epsilon_0 A$).

$$V = \frac{Q}{\epsilon_0 A}\left(d - t - \frac{t}{\epsilon_r}\right) \tag{2.73}$$

Notice that $V < V_0$ where $V_0 = E_0 d = Qd/\epsilon_0 A$. This decrease in V results in an increase in capacitance C which is given by

$$C = \frac{Q}{V} = \frac{\epsilon_0 A}{d + t\left(\frac{1 - \epsilon_r}{\epsilon_r}\right)}$$

or

$$C = \frac{\epsilon_r C_0}{\frac{t}{d} + \epsilon_r\left(1 - \frac{t}{d}\right)} \tag{2.74}$$

where $C_0 = \epsilon_0 A/d$ is the vacuum value. If the dielectric fills the entire space between the plates, i.e. if $t = d$, then we have [setting $t = d$ in Eq (2.74)],

$$C = \epsilon_r C_0 \tag{2.75}$$

Thus, if the space between the plates of a parallel-plate capacitor is completely filled with a dielectric of dielectric constant ϵ_r, its capacitance is increased by a factor ϵ_r compared to the vacuum value. Incidentally this is how ϵ_r is actually defined and measured.

Note: Although the *effects* with conducting and dielectric slabs are qualitatively similar (they both increase the capacitance) their *causes* are very different: it is the large-scale movement of free charges in the former (conductor) and the small-scale alignment of dipoles (i.e. polarization) in the latter (dielectric).

Electric Displacement

While dealing with problems on the behaviour of dielectrics in an electric field, especially the propagation of electromagnetic waves in dielectrics, a particular linear combination

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P} \tag{2.76}$$

simplifies the analysis. This vector **D** is called the *electric displacement.* The use of **D** is convenient, not necessary. All the physics of dielectrics could be discussed without ever using the displacement vector **D**. It is of historical interest and also has some practical advantages. **D** is related to **E** as

$$\mathbf{D} = \epsilon \mathbf{E} \tag{2.77}$$

where ϵ is the electric permittivity of the dielectric. Since $\mathbf{P} = \epsilon_0 \chi \mathbf{E}$, Eq. (2.76) becomes

$$\epsilon \mathbf{E} = \epsilon_0 \mathbf{E} + \epsilon_0 \chi \mathbf{E}$$

giving

$$\epsilon_r = \frac{\epsilon}{\epsilon_0} = 1 + \chi$$

which is Eq. (2.72).

EXAMPLE 2.24 A spherical capacitor has an inner sphere of radius 12 cm and an outer sphere of radius 13 cm. The outer sphere is earthed and the inner sphere is given a charge of 2.5 μC. The space between the concentric spheres is filled with a liquid of dielectric constant 32. (a) Determine the capacitance of the capacitor. (b) What is the potential of the inner sphere? (c) Compare the capacitance of this capacitor with that of an isolated sphere of radius 12 cm. Explain why the latter is much smaller.

Solution:

Radius of outer sphere (r_b) = 13 cm = 13×10^{-2} m
Radius of inner sphere (r_a) = 12 cm = 12×10^{-2} m
Charge on inner sphere (Q) = $+2.5 \times 10^{-6}$ C
Dielectric constant (ϵ_r) = 32

The charge $+Q$ induces a charge $-Q$ on the outer sphere. The potential V_a of the inner sphere = potential due to $+Q$ on inner sphere + potential due to $-Q$ on outer sphere

or

$$V_a = \frac{Q}{4\pi\epsilon r_a} - \frac{Q}{4\pi\epsilon r_b} \tag{i}$$

where ϵ = permittivity of dielectric = 32.

Since the outer sphere is earthed, its potential $V_b = 0$. Therefore.

Potential difference

$$V = V_a - V_b$$

$$= \frac{Q}{4\pi\epsilon_r\epsilon_0}\left(\frac{1}{r_a} - \frac{1}{r_b}\right) \tag{ii}$$

$$\therefore \text{ Capacitance } C = \frac{Q}{V} = \frac{4\pi\epsilon_r\epsilon_0 r_a r_b}{(r_b - r_a)} \tag{iii}$$

(a) Substituting the values of r_a, r_b and ϵ_0 in (iii) and solving we get

$$C = 5.5 \times 10^{-9} \text{ F}$$

(b)
$$V_a = \frac{Q}{4\pi\epsilon_r\epsilon_0}\left(\frac{1}{r_a} - \frac{1}{r_b}\right)$$

Substituting the values of all the quantities and solving, we get

$$V_a = 4.5 \times 10^2 \text{ V}$$

(c) The potential of an isolated sphere of radius r_a is

$$V_0 = \frac{Q}{4\pi\epsilon_0 r_a} \qquad \text{(iv)}$$

$$\therefore \quad C_0 = \frac{Q}{V_0} = 4\pi\epsilon_0 r_a \qquad \text{(v)}$$

Comparing (iv) with (ii) we notice that $V_0 > V$. The reason is that the outer sphere has been effectively removed to infinity ($r_b \to \infty$); hence the potential due to induced charge is absent. Since $V_0 > V$, $C_0 < C$. Substituting for ϵ_0 and r_a in (v) gives

$$C_0 = 1.3 \times 10^{-11} \text{ F}$$

EXAMPLE 2.25 A parallel-plate capacitor is to be designed with a voltage rating 1 kV, using a material of dielectric constant 3 and dielectric strength 10^7 Vm^{-1}. If the field is not to exceed 10% of the dielectric strength, what minimum area of the plate is required to have a capacitance of 50 pF?

Solution: The dielectric strength of a material is the maximum electric field it can tolerate without breakdown. Electric fields higher than this maximum limit will produce partial ionization of the material, and as a result the material begins to conduct charge. Thus

$$E_{\text{max}} = 10\% \text{ of } 10^7 \text{ Vm}^{-1} = 10^6 \text{ Vm}^{-1}$$

Now
$$E = \frac{Q}{A\epsilon} = \frac{Q}{A\epsilon_r\epsilon_0}$$

But
$$Q = CV$$

$$\therefore \quad E = \frac{CV}{A\epsilon_r\epsilon_0}$$

Thus
$$A_{\text{min}} = \frac{CV}{\epsilon_r\epsilon_0 E_{\text{max}}} \qquad \text{(i)}$$

It is given that $C = 50\ \text{pF} = 50 \times 10^{-12}\ \text{F}$, $V = 1\ \text{kV} = 10^3\ \text{V}$, $\epsilon_r = 3$.

$$\therefore \qquad A_{\min} = \frac{(50 \times 10^{-12}) \times 10^3}{3 \times (8.85 \times 10^{-12}) \times 10^6}$$

$$= 1.9 \times 10^{-3}\ \text{m}^2 = 19\ \text{cm}^2$$

2.10 ATMOSPHERIC ELECTRICITY

The earth's atmosphere extends to about 300 km above its surface. Depending on the gross physical features, the earth's atmosphere is divided into the following four layers:

(*a*) *Troposphere*: The layer of the atmosphere between the surface of the earth and an altitude of about 12 km is called the *troposphere*. The density of air in this layer falls from $\rho_0(\sim 1.29\ \text{kg m}^{-3})$ at ground level to about $\rho_0/10$ at the end of troposphere. The temperature falls from about 15°C (at ground level) to about −50°C (at 12 km altitude).

(*b*) *Stratosphere*: The layer of the atmosphere between altitudes 12 and 50 km is called the *stratosphere*. In this layer, the density decreases from $\rho_0/10$ to $\rho_0/10^3$ and the temperature rises from about −50°C to about 10°C.

(*c*) *Mesosphere*: The layer of the atmosphere between altitudes 50 and 80 km is called the *mesosphere*. In this layer, the density decreases from $\rho_0/10^3$ to $\rho_0/10^5$ and the temperature falls from 10°C to about −90°C.

(*d*) *Ionosphere*: The ionosphere extends from 80 km to about 300 km above the surface of the earth. The density of air in this layer falls from $\rho_0/10^5$ to $\rho_0/10^{10}$ and the temperature rises from −90°C to about +400°C.

The densities and temperatures given above are approximate average values; there are obviously no sharp boundaries between the different layers.

Electrical Conductivity of the Atmosphere

The atmosphere near the surface of the earth is a very poor conductor of electricity. The electrical conductivity of the atmosphere increases quite dramatically with altitude. The troposphere is a poor conductor whereas the ionosphere is a very good conductor of electricity. The conductivity of the atmosphere is mainly due to the ionization of its molecules. The fact that ions are present in air has been known for a long time because it was observed that the leaves of a charged electroscope gradually collapse when left in air.

What causes the ionization of air? For a long time, physicists believed that the ionization of air was caused by radioactive materials present inside the earth's crust. In 1910, Gockel and Hess, independently carried out measurements of ionization at different altitudes by carrying instruments (such as electroscopes) in balloons and found that ionization increases with altitude. At an altitude of about 8 km, the ionization was found to be about 10 times as great as at sea level. These observations cannot be explained on the assumption that the source of the ionization is inside the earth. The observed results can be explained if we assume that radiations of very high penetrating power enter our atmosphere from above, i.e. the source of the ionization lies outside the earth. These radiations were named *cosmic rays*. Cosmic rays are charged particles with very high energies of the order of 10^{14} MeV. When these rays pass through the upper atmosphere of the earth, they collide violently (due to their high energy) with the molecules of the atmosphere and ionize them, i.e. remove electrons from them. The maximum degree of ionization occurs in the topmost layer (the ionosphere) which is, therefore, the most conducting layer. The layer nearest the earth (the troposphere) in the least conducting.

Note: The origin of cosmic rays is not well understood yet. Earlier it was believed that these rays originated from the sun, but this idea was soon abandoned because the particles ejected from the sun cannot have energies as high as 10^{14} MeV. It is, therefore, believed that cosmic rays originate from other sources within our galaxy or even beyond it. Recent observations have shown that these rays are produced by the explosion of certain stars and even galaxies.

Main Average Electrical Properties of Atmosphere

We now list the main average electrical properties and phenomena associated with them in the 50 km thick blanket enveloping the earth. These properties were found through a series of experimental observations.

1. At ground level, there exists an electric field with a strength of about 100 Vm^{-1} directed vertically downward all over the earth. The strength of the field decreases gradually with altitude; at 10 km it is quite small, and at 50 km it is negligible.
2. Between the conducting layer at the top of the stratosphere

(about 50 km above the earth) and the surface of the earth, there is a potential drop of about 4×10^5 V; most of the potential drop occurring at low altitudes where the electric field is stronger.

3. Because the earth is a good conductor (and hence an equipotential surface), the downward electric field at ground level produces a negative surface density of charge all over the earth. The charge density is of the order of -10^{-9} C m^{-2}. Thus the total negative charge on the whole of the earth's surface is

$$Q = 4\pi r_a^2\sigma \simeq -10^6 \text{ C}$$

about a million coulombs.

4. The downward electric field at ground level makes the positive ions move downward and the negative ions upward. The flow of charges gives rise to a current flowing into the earth. The current density (i.e. current per unit area) has an average value of $J = 3\times 10^{-12}$ C s^{-1} m^{-2} (or A m^{-2}). Since the total surface area of the earth is $4\pi r_a^2 = 4\pi\times(6.4\times 10^6)^2 \simeq 6.8\times 10^{14}$ m^2 this would mean that about +1800 C of charge is being deposited on the earth in each second.

Thunderstorms and Lightning Flashes

We have mentioned that the total charge on the earth is about -10^6C and that it receives about +1800 C of charge every second. This means that the entire negative charge on the earth will be neutralized in about 15 minutes. But this does not happen; the reason is that the facts listed above describe the *steady state* of the earth and the atmosphere. What effects are responsible for maintaining the potential difference of 4×10^5 V between the stratosphere and the earth and the negative charge of 10^6 C on the earth? These are thunderstorms and lightning flashes which occur all over the earth, especially during the monsoon. How this happens is not very well understood yet. We list below a few salient features which are mainly the result of experimental observations.

(a) On an average, about 4×10^4 thunderstorms occur per day in the whole world which means that a storm starts off somewhere every two seconds. Each storm lasts roughly for one hour.

(b) By a rather complex process, there occurs a separation between charged ions in a thundercloud; the positive charge is carried upward to a height of about 6 km and the negative charge collects at about 2 to 3 km above the ground (the bottom of the cloud).

(c) The total amount of the negative charge is as high as about -20 C which creates a potential difference of about 10^8 V between the bottom of the cloud and the earth (Fig. 2.48); the latter being at a higher potential. The separation between positive and negative charges is about 3 km. Consequently, an electric field of strength about 10^4 Vm^{-1} is created directed upward.

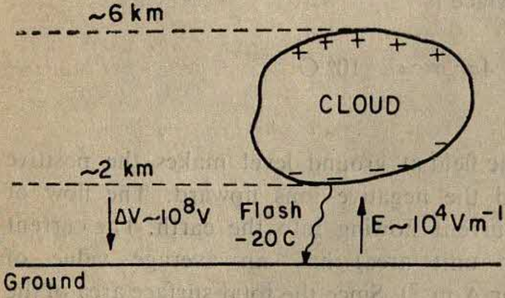


Fig. 2.48 *Mechanism of lightning flashes*

(d) Such high fields cause the *electrical breakdown* of the intervening air, i.e. it is ionized and hence begins to conduct charge. Thus, a huge amount of charge is conducted to the earth in the form of lightning flashes. Each flash deposits about -20 C of charge on the earth.

(e) After each flash, the cloud becomes charged again and is then ready for the next flash. So with about 90 flashes per second, the thunderstorms pump about -1800 C of charge into the earth per second. This is exactly what is needed to counterbalance the $+1800$ C of charge per second in the regions of the earth which do not experience any significant thunderstorms.

2.11 THE ELECTROSTATIC GENERATOR

We have seen how large charge concentrations leading to high potential differences and fields are created in nature in thunderstorms. Man has also built huge machines (called electrostatic generators) which can produce potential differences of the order of several millions of volts. The chief application of such large machines is the use of this high potential difference to accelerate charged particles to high energies of the order of a few million electron-volts (MeV). These high energy particles are used in 'atom splitting' experiments.

The electrostatic generator was first conceived by Lord Kelvin in 1890 and put into useful practice in its modern form by RJ Van de Graaff in 1931.

Principle of Van de Graaff Generator

We have learnt that a spherical conducting shell of radius R has a capacitance $C = 4\pi\epsilon_0 R$ (see section 2.8). Since $V = Q/C = Q/(4\pi\epsilon_0 R)$, it follows that, for a given change Q, the voltage V will be high if the radius R is small. We also know that at sharply pointed surfaces of a charged conductor, the electric field becomes very large compared to smooth parts of the surface. These facts form the basis of a generator,

An electrostatic generator essentially separates charges by induction and accumulates them at sharp points. Figure 2.49 shows the essential parts of one such generator, made by Van de Graaff. It consists of a hollow metal sphere S supported on an insulating tube T many metres high. A strong silk belt B runs over pulleys P_1 and P_2, the lower pulley P_2 being continuously driven by a motor. Two electrodes E_1 and E_2 which are sharply pointed combs pass close to the belt near its top and bottom as shown. The electrode E_1 is given a positive potential with respect to the earth by a battery.

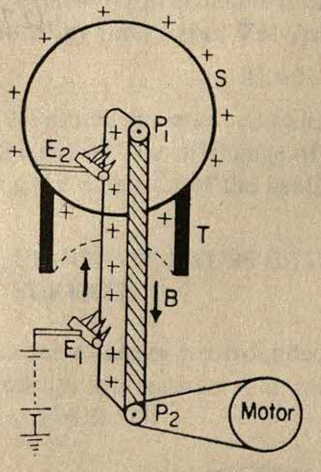


Fig. 2.49 *Principle of Van de Graaff generator*

The high electric field at the points ionizes the air near them and so positive charges are repelled on to the belt which carries this charge up to the sphere. Here it induces a negative charge on the points of electrode E_2 and a positive charge on the sphere which is connected to the blunt end of E_2. The high electric field near the points of E_2 ionizes the air there. The negative charges produced during ionization are repelled to the belt, thus neutralizing its positive charge before it passes over the pulley. In this way, positive charge keeps accumulating on the sphere until its potential is about a million volts relative to the earth.

SUPPLEMENTS

S.1 GENERALIZATION OF RELATION (2.39) TO THREE DIMENSIONS

We know that the potential difference between two points in an electric field **E** is related to the line integral of **E** as

$$d\phi = -\mathbf{E}\cdot d\mathbf{l}$$

If θ is the angle betwee **E** and d**l**, we can write this relation as

$$d\phi = -E\,dl\cos\theta$$

or

$$E\cos\theta = -\frac{\partial\phi}{\partial l} \tag{2.78}$$

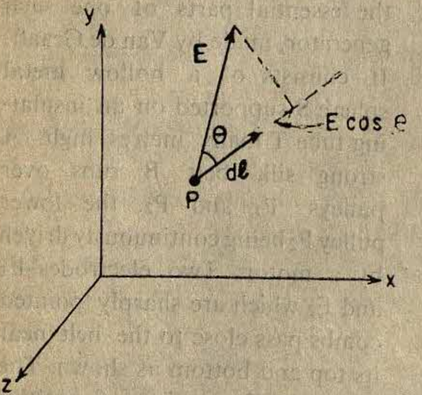


Fig. 2.50 *E cos θ is the component of* E *along dl*

As shown in Fig. 2.50, $E\cos\theta$ is the component of the electric field vector **E** in the direction of the vector displacement d**l**. Suppose we take d**l** along the x-direction, then $E\cos\theta$ will be the component of **E** along the x-direction. We call it E_x. Equation (2.78) then reduces to

$$E_x = -\frac{\partial\phi}{\partial x} \tag{2.79}$$

which is our Eq. (2.39).

Now vector **E** can be written in terms of its three rectangular components E_x, E_y and E_z along x, y and z axes as

$$\mathbf{E} = \hat{\mathbf{x}}E_x + \hat{\mathbf{y}}E_y + \hat{\mathbf{z}}E_z$$

where $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$ and $\hat{\mathbf{z}}$ are the unit vectors along x, y and z axes. But each component of **E** can be written through Eq. (2.78) as

$$E_x = -\frac{\partial \phi}{\partial x}$$

$$E_y = -\frac{\partial \phi}{\partial y}$$

$$E_z = -\frac{\partial \phi}{\partial z}$$

or

$$\mathbf{E} = -\left(\hat{\mathbf{x}}\frac{\partial \phi}{\partial x} + \hat{\mathbf{y}}\frac{\partial \phi}{\partial y} + \hat{\mathbf{z}}\frac{\partial \phi}{\partial z}\right)$$

The expression in brackets is called the *gradient* of ϕ and is usually denoted by the symbol $\nabla\phi$. Thus

$$\mathbf{E} = -\nabla\phi \tag{2.80}$$

By definition, the gradient (which means slope) is the maximum value of the spatial rate of change of ϕ and is obtained by evaluating $\partial\phi/\partial l$ along the direction of the field vector itself.

S.2 PROOF OF GAUSS'S THEOREM FOR A GENERAL SURFACE

Consider a charge q surrounded by a general closed surface S. For simplicity, we assume that each radial line from q cuts S in only one point (Fig. 2.51).

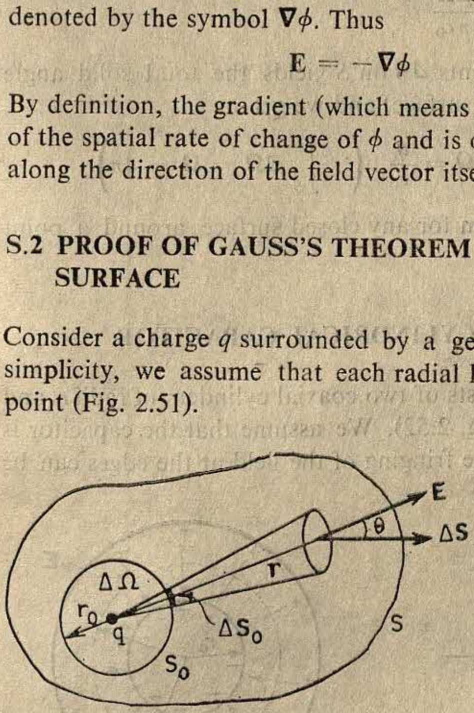


Fig. 2.51 *Gauss's theorem for a general closed surface around a point charge*

To prove Gauss's theorem for a general closed surface, draw around q a sphere S_0 of radius r_0 so that S_0 is completely enclosed within S. If $\Delta\mathbf{S}$ is a small area element on S around a point **r** on S, draw a small cone of radial lines from q to the boundary of $\Delta\mathbf{S}$. These lines will cut out a small patch ΔS_0 on S_0 as shown, which will point radially outwards. Now from geometry, we find that

(a) $\Delta \mathbf{S}$ and $\Delta \mathbf{S}_0$ subtend the same *solid angle* $\Delta\Omega$ at q,

(b) $\Delta \mathbf{S}$ makes some angle θ with the radial direction $\hat{\mathbf{r}}$, and

(c) $|\Delta \mathbf{S}_0| = r_0^2 \Delta\Omega$ and it is the projection $(\Delta\mathbf{S}\cdot\hat{\mathbf{r}})$ of $\Delta\mathbf{S}$ along $\hat{\mathbf{r}}$.

Since the field is radial, we have

$$\mathbf{E} = \frac{q}{4\pi\epsilon_0}\cdot\frac{\hat{\mathbf{r}}}{r^2}$$

Also $$\Delta\mathbf{S}\cdot\hat{\mathbf{r}} = r^2\Delta\Omega$$

Therefore $$\mathbf{E}\cdot\Delta\mathbf{S} = \frac{q\Delta\Omega}{4\pi\epsilon_0}$$

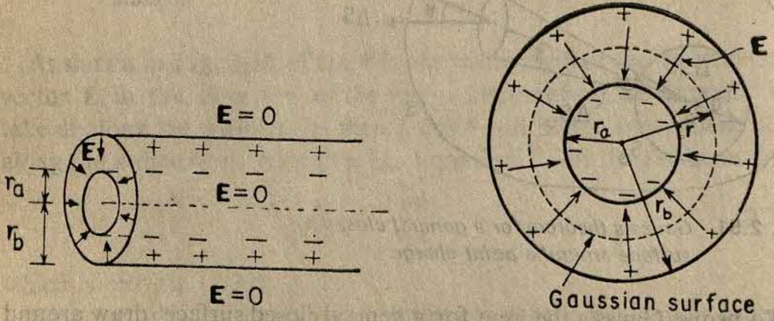
Now summing over all elements $\Delta\mathbf{S}$ on S yields the total solid angle of $\Omega = 4\pi$ (as in a spherical surface) and we get

$$\int_S \mathbf{E}\cdot\Delta\mathbf{S} = \frac{q}{4\pi\epsilon_0}\int_S \Delta\Omega = \frac{q}{\epsilon_0} \quad \left(\because \int_S \Delta\Omega = \Omega = 4\pi\right)$$

which proves Gauss's theorem for any closed surface around a point charge.

S.3 CAPACITANCE OF A CYLINDRICAL CAPACITOR

A cylindrical capacitor consists of two coaxial cylinders of radii r_a and r_b ($> r_a$) and length L (Fig. 2.52). We assume that the capacitor is very long ($L \gg r_b$) so that the fringing of the field at the edges can be ignored.



(a) (b)

Fig. 2.52 *(a) A cylindrical capacitor (b) A cross-section of the capacitor*

To calculate the capacitance, we draw a convenient Gaussian surface which is a cylinder of radius, say r, (shown by a dashed circle) and length L, closed at the ends by plane caps. If Q is the total charge enclosed in this surface, we have from Gauss's theorem:

$$\int_S \mathbf{E}\cdot d\mathbf{S} = \frac{Q}{\epsilon_0}$$

Since $\mathbf{E}$ is radial, the surface integral is just $E(2\pi rL)$ so that

$$E(2\pi rL) = \frac{Q}{\epsilon_0}$$

or

$$E = \frac{Q}{2\pi\epsilon_0 rL} \tag{2.81}$$

The potential difference V, by definition, is

$$V = -\int_{r_a}^{r_b} \mathbf{E}\cdot d\mathbf{l}$$

Note that $\mathbf{E}$ and $d\mathbf{l}$ ($= dr$) point in opposite directions. Hence

$$V = \int_{r_a}^{r_b} E\, dr \tag{2.82}$$

Using Eq. (2.81) in Eq. (2.82) gives

$$\begin{aligned} V &= \int_{r_a}^{r_b} \frac{Q}{2\pi\epsilon_0 L}\cdot\frac{dr}{r} \\ &= \frac{Q}{2\pi\epsilon_0 L}\int_{r_a}^{r_b}\frac{dr}{r} \\ &= \frac{Q}{2\pi\epsilon_0 L}\ln\left(\frac{r_b}{r_a}\right) \end{aligned}$$

$$\therefore \qquad \text{Capacitance } C = \frac{Q}{V} = \frac{2\pi\epsilon_0 L}{\ln(r_b/r_a)} \tag{2.83}$$

As in the case of a parallel-plate capacitor, the capacitance of a cylindrical capacitor depends only on the geometry of the capacitor.

SUMMARY

The line integral of any general electrostatic field is path-independent, depending only on the end points of the path. The line integral of an electric field around a closed path is zero. The line integral of an electrostatic field between any two points equals the negative of the potential difference between them. The potential difference between two points in an electrostatic field is the amount of work done per unit charge to move it from one point to the other against the electrostatic force of the field along any path. The net work done in carrying a charge over a closed loop in an electrostatic field is zero. The electric potential at a point in an electrostatic field is the work done per unit charge to bring a small charge in from infinity to that point along any path.

The electric potential energy of a system of point charges is the amount of work done to assemble the system of charges by bringing them in from an infinite distance.

The electron-volt is the potential energy gained or lost by an electron in moving through a potential difference of 1 volt. 1 eV$=1.6\times10^{-19}$J.

The electric field is equal to the negative of the potential gradient, i.e.,

$$\mathbf{E} = -\nabla\phi$$

Any two-dimensional surface over which the electric potential is constant is called an equipotential surface. The electric field is always at right angles to an equipotential surface. No work is required to move a charge along an equipotential surface.

Gauss's theorem states that the surface integral of the electrostatic field due to any sources is proportional to the total charge contained within the closed surface, i.e,

$$\int_S \mathbf{E}\cdot d\mathbf{S} = \frac{q}{\epsilon_0}$$

Materials which have a large number of free charges are called conductors. There is no net electric field or net charge in the interior of a conductor; charge resides on its surface. The surface charge density is high where the radius of curvature is small. The surface of a conductor is an equipotential surface. Just outside a conductor, the electric field is perpendicular to the surface.

Materials which have only bound charges are called insulators or dielectrics. Polar dielectrics have a permanent dipole moment. Nonpolar dielectrics acquire an induced dipole moment when placed in an electric field.

A capacitor is any combination of two conducting bodies, of any shape and size, in certain definite positions in space relative to each other. Capacitors are used for storing charge. The capacitance C of a capacitor is a measure of how much charge Q the capacitor can store at a given voltage V. In fact $Q = CV$. The capacitance of a capacitor is increased by introducing a conducting or insulating material in the space between its two conducting plates. The factor by which the capacitance is multiplied by the introduction of a dielectric is called the dielectric constant of the latter.

The two vectors that characterize the behaviour of a dielectric in an external electric field are electric polarization and electric displacement vectors.

EXERCISES

A. Short-Answer Questions

1. What is the line integral of an electrostatic field?
2. Define electric potential and potential difference.
3. What do you understand by electric potential energy? State its SI unit.
4. What is an electron-volt? How is it related to a joule?
5. Define an equipotential surface.
6. State Gauss's theorem in electrostatics.
7. Distinguish between conductors and insulators. Give two examples of each.
8. What are polar and non-polar dielectrics?
9. Define capacitance. State and define its SI unit.
10. What is dielectric constant?
11. Define electric polarization and electric susceptibility.
12. How is electric susceptibility related to dielectric constant?
13. What will happen to a person on an insulated stand if his potential is increased to say, 10000 V?
14. Electrons tend to move towards the region of a higher potential. Is it true or false? Give the reason for your answer.
15. Can two different equipotential surface intersect? Explain.
16. If electric field E is zero at a point, the electric potential ϕ must also be zero at that point. Is this statement true? Give the reason for your answer.
17. If you know E at a point, can you calculate ϕ at that point? If not, what further information do you need?
18. If ϕ is constant throughout a given region of space, what can you say about E in that region?
19. How can you ensure that ϕ in a given region of space will have a constant value?
20. An uncharged metal sphere is suspended by a silk thread and placed in a uniform electric field. What can you say about the electric field inside the sphere? Will your answer change if the sphere is given a charge?
21. Can there be a potential difference between two adjacent conductors that carry equal positive charges?

B. Long-Answer Questions

1. What do you understand by the line integral of an electrostatic field? Evaluate the integral for the case of the electric field of a static point charge and hence show that it is path-independent.
2. Show that the line integral of an electrostatic field is zero around a closed path.

3. How does the use of the line integral lead us to the concept of electric potential?
4. Deduce the relationship between the line integral and the work done to move a charge from one point to another in an electrostatic field. Hence define potential and potential difference.
5. What are conservative and non-conservative fields? Show that the electrostatic field is conservative.
6. Obtain an expression for electric potential due to: (a) a point charge, (b) two point charges and (c) N point charges.
7. Obtain an expression for the potential at a point P distant r from the centre O of a dipole, the direction OP making an angle θ with dipole moment **p**. Show that the potential at a point on the perpendicular bisector of a dipole is zero.
8. What is the potential energy of two point charges q_1 and q_2 separated by a distance r in vacuum? How will you generalize to N point charges?
9. Deduce the relation between electric field and potential gradient in the one-dimensional case. How can you generalize to three dimensions?
10. What are equipotential surfaces? Show that the electric field is always at right angles to an equipotential surface.
11. State Gauss's theorem and prove it for the case of an isolated point charge.
12. Use Gauss's theorem to find the electric field due to: (a) a line of charge, (b) a thin sheet of charge, (c) a thin spherical shell of charge and (d) a solid sphere of charge.
13. A point charge is placed at the centre of a spherical Gaussian surface. Will the surface integral change if (a) the surface is replaced by a cube of half the volume, (b) the charge is moved off-centre of the original sphere, still remaining inside, (c) the charge is moved just outside the original sphere, (d) a second charge is placed near and outside the original sphere, and (e) a second charge is placed inside the original sphere?
14. A surface encloses an electric dipole. What can you say about the surface integral?
15. A spherical rubber balloon carries a charge that is uniformly distributed over its surface. The balloon is inflated more by blowing. How does the electric field vary for points: (a) inside, (b) at the surface, and (c) outside the balloon?
16. State and explain four electrical properties of a conductor.
17. What is electrostatic shielding? How is it practically useful?
18. What are polar and non-polar dielectrics? Define the terms atomic polarizability, electric polarization, electric susceptibility and dielectric constant.
19. Describe qualitatively what happens if a dielectric is placed in a uniform electric field.
20. What do you understand by the term 'capacitance'? State and define its SI unit. How large is this unit?
21. Deduce the expression for the capacitance of: (a) a parallel-plate capacitor, and (b) a cylindrical capacitor. Neglect edge effects.
22. What is a spherical capacitor? Obtain an expression for its capacitance. Hence find the capacitance of an isolated charged sphere.
23. A capacitor is connected across a battery. Why does each plate receive a charge of exactly the same magnitude? Is this true even if the plates have different areas?

24. Deduce the expression for the capacitance of a parallel-plate capacitor when: (a) a conducting slab, and (b) a dielectric slab, are inserted between the plates. Assume the slab thickness to be less than the plate separation. Discuss similarities and differences between the two situations.
25. For a given potential difference, does a capacitor store more or less charge with a dielectric than it does without a dielectric (vacuum)? Give in detail the reason for your answer.
26. A capacitor is charged by using a battery, which is then disconnected. A dielectric slab is then inserted between the plates. Describe qualitatively what happens to the charge, the capacitance, the potential difference and the electric field strength.
27. What would be your answer to Question 26 if the battery were not disconnected?
28. Three different capacitors are connected: (a) in series, and (b) in parallel. Find the equivalent capacitance in each case.
29. Briefly describe the main average electrical properties of the atmosphere from ground level up to the stratosphere. What is the mechanism of lightning flashes?
30. Describe, with the aid of a labelled diagram, a Van de Graaff generator, explaining the physical principles of its action.

C. Multiple-Choice Questions

1. The SI unit of the line integral of an electric field is

(a) V
(b) Vm^{-1}
(c) C
(d) NC^{-1}

2. What is the SI unit of the surface integral of an electric field?

(a) V
(b) Vm^{-2}
(c) V m
(d) $C\,m^{-2}$

3. What is the potential energy of two equal positive point charges $1\mu C$ each held 1 m apart in air?

(a) 1 J
(b) 1 eV
(c) 9×10^{-3} J
(d) zero

4. Which of the following statements is *not* true for an equipotential surface?

(a) The potential difference between any two points on the surface is zero.
(b) The electric field is always at right angles to the surface.
(c) No work is done to move a charge along the surface.
(d) Equipotential surfaces are always spherical.

5. A hollow metal sphere of radius 10 cm is charged to a potential of 100 V. The potential at the centre of the sphere is

(a) zero
(b) 100 V
(c) the same as that at a point 10 cm away from the surface
(d) the same as that at a point 50 cm away from the surface

6. Two equal negative charges $-q$ are placed at points $(0, a)$ and $(0, -a)$ on

the y-axis. A positive charge Q is released from rest at a point $(2a, 0)$ on the x-axis. The charge Q will

(a) execute simple harmonic motion about the origin
(b) move to the origin and remain at rest there
(c) move to infinity
(d) execute oscillatory but not simple harmonic motion

7. Four capacitors each of capacitance 50 μF are connected as shown in Fig. 2.53. The dc voltmeter reads 100 V. The magnitude of charge on each plate of each capacitor is

(a) 2×10^{-3} C
(b) 5×10^{-3} C
(c) 0.2 C
(d) 0.5 C

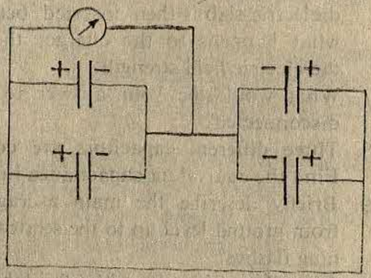


Fig. 2.53

8. An air-filled parallel-plate capacitor has a capacitance of 1 pF. The plate separation is doubled and the space between them is completely filled with wax. If the capacitance increases to 2 pF, what is the dielectric constant of wax?

(a) 2.0 (b) 4.0
(c) 6.0 (d) 8.0

9. The work done in carrying a charge q once round a circle of radius r with a charge Q at the centre is

(a) $\frac{qQ}{4\pi\epsilon_0 r}$ (b) $\frac{qQ}{4\pi\epsilon_0 r^2}$
(c) $\frac{qQ}{4\pi\epsilon_0}\left(\frac{1}{2\pi r}\right)$ (d) zero

10. When a capacitor is connected to a battery:

(a) no current flows in the circuit at all
(b) a current flows in the circuit for some time, after which it decreases to zero
(c) the current keeps on increasing, reaching a maximum value when the capacitor has been charged to the voltage of the battery
(d) an alternating current flows in the circuit

11. A hollow metal sphere placed on an insulating stand is charged positively. The electric potential inside the sphere is

(a) zero
(b) the same as on the surface
(c) greater than the potential on the surface
(d) less than the potential on the surface

12. Three capacitors of capacitances 8 μF, 12 μF and 24 μF are connected (a) in series and then (b) in parallel. What is the ratio of the equivalent capacitance in case (a) to that in case (b)?

(a) 1 : 11 (b) 11 : 1
(b) 1 : 1 (d) 1 : 3

13. Two identical parallel-plate capacitors are connected in parallel. The combination is connected to 6 V battery. The battery is now disconnected and the two capacitors are joined in series (Fig. 2.54). What is the potential difference between points A and B?

(a) 3 V
(b) 6 V
(c) 9 V
(d) 12 V

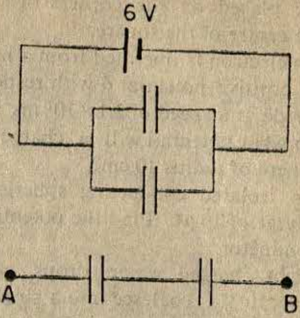


Fig. 2.54

14. Two identical metal plates, separated by a distance d, form a parallel-plate capacitor. A metal sheet of thickness $d/2$, of the same area as that of either plate, is inserted between the plates. The ratio of the capacitance after the insertion of the sheet to that before insertion is

(a) $\sqrt{2} : 1$ (b) $2 : 1$
(c) $1 : 1$ (d) $1 : \sqrt{2}$

15. Four capacitors are connected as shown in Fig. 2.55. The equivalent capacitance between points P and Q is

(a) 4 μF (b) $\frac{1}{4}$ μF

(c) $\frac{3}{4}$ μF (d) $\frac{4}{3}$ μF

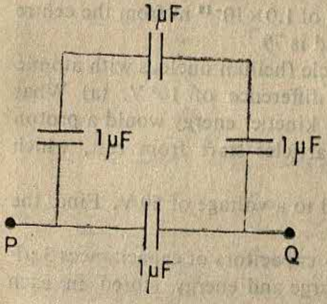


Fig. 2.55

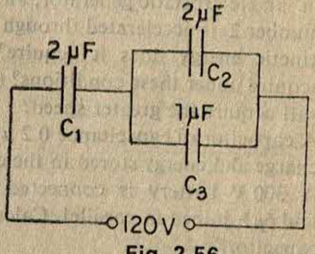


Fig. 2.56

16. Three capacitors C_1, C_2 and C_3 are connected to a supply of 120 V as shown in Fig. 2.56. What is the potential difference across C_1?

(a) 48 V (b) 60 V
(c) 72 V (d) 90 V

D. Numerical Problems

(Take $\epsilon_0 = 8.85 \times 10^{-12}$ Fm^{-1}, $e = 1.6 \times 10^{-19}$ C wherever necessary.)

1. Calculate the electric potential at the surface of the nucleus of a gold atom. The radius of a gold nucleus is 6.6×10^{-15} m and the atomic number of gold is 79.

2. Four point charges $q_1 = 1.0\,\mu C$, $q_2 = 2.0\,\mu C$, $q_3 = -3.0\,\mu C$ and $q_4 = 5.0\,\mu C$ are placed on the corners of a square of side 1.0 m. Find the potential at the centre of the square.
3. An electron is liberated from a hot cathode and attracted by an anode kept at positive potential ϕ with respect to the cathode. If the electron strikes the anode at a speed of $2.1 \times 10^7\ ms^{-1}$, find the value of ϕ.
4. To what potential will a charge of 10^{-8} C raise an insulated conducting sphere of radius 10 cm?
5. An isolated conducting spherical shell of radius 10 cm carries a positive charge of 10 μC. Find the potential and electric field at the surface of the conductor.
6. Eight charged water drops, each of radius 1.0 mm and a charge of 1.0×10^{-10} C coalesce into a single drop. Find the potential of the big drop.
7. A tiny nucleus has a charge $+ 50e$. Find the potential at a distance of 1.0×10^{-12} m from the nucleus. If a proton (mass $= 1.67 \times 10^{-27}$ kg) is released from rest at this point, how fast will it be moving when it is 1.0 m from the nucleus?
8. A small sphere of mass 2.0×10^{-4} kg hangs vertically by a thread between two vertical plates 5.0 cm apart.The charge on the sphere is 6.0×10^{-9} C. What potential difference must be applied between the plates so that the thread is inclined at 10° with the vertical?
9. Calculate the potential energy in eV of the electron-proton system of a hydrogen atom. The radius of the electron orbit is 5.29×10^{-11} m.
10. The mutual electric potential energy between two protons in a uranium nucleus is 2.4×10^5 eV. What is the distance between the protons?
11. Three point charges 1 μC, 2 μC and −4 μC are placed at the vertices of an equilateral triangle of side 10 cm. Find potential energy of the system.
12. What is the potential gradient at a distance of 1.0×10^{-12} m from the centre of a gold nucleus? The atomic number of gold is 79.
13. In an electrostatic generator, an alpha particle (helium nucleus with atomic number 2) is accelerated through a potential difference of 10^6 V. (a) What kinetic energy does it acquire? (b) What kinetic energy would a proton acquire under these conditions? (c) If both particles start from rest, which will acquire the greater speed?
14. A capacitor of capacitance 0.2 μF is charged to a voltage of 50 V. Find the charge and energy stored in the capacitor.
15. A 300 V battery is connected across two capacitors of capacitances 3 μF and 6μF joined in parallel. Calculate the charge and energy stored in each capacitor.
16. What would be the charge and energy stored in each capacitor of Question 15 if they were joined in series?
17. A 300 V battery is connected across a 2.0 μF capacitor and an 8.0 μF capacitor joined in series. When the capacitors are fully charged, the battery is disconnected. The charged capacitors are then reconnected with like plates joined together. What is the charge and potential difference for each capacitor?
18. What would be the charge and potential difference for each capacitor in Question 17 if the charged capacitors were connected with plates of opposite sign joined together?

19. A 100 pF capacitor is charged by a 50 V battery. The battery is then disconnected. The charged capacitor is now connected in parallel with an uncharged second capacitor. The potential difference is found to drop to 35 V. Find the capacitance of the second capacitor.
20. A 4 μF capacitor is charged to a voltage of 100 V and a 6 μF capacitor is charged to a voltage of 200 V. They are now joined with plates of like charges connected together. Find: (a) potential difference across each after joining, (b) the total energy stored before joining, and (c) the total energy stored after joining.
21. A parallel-plate capacitor with plate area 200 cm^2 and plate separation 5.0 mm is connected to a 500 V supply. (a) Find the charge, energy stored and electric field strength between the plates. (b) The space between the plate is completely filled with oil. It is found that the capacitor now collects an additional charge of 2.65×10^{-8} C. What is the dielectric constant of the oil?
22. A capacitor is fully charged by a 200 V supply. It is then discharged through a small coil of wire embedded in a thermally insulated block of specific heat 2.5×10^2 J kg^{-1} K^{-1} and of mass 0.1 kg. The temperature of the block is found to rise by 0.4 K. Calculate the capacitance of the capacitor.
23. A charge Q is distributed over two concentric hollow spheres of radii r and R ($>r$) such that the surface charge densities are equal. Find the potential at the common centre.
24. Four capacitors $C_1 = 1.0$ μF, $C_2 = 2.0$ μF, $C_3 = 3.0$ μF and $C_4 = 4.0$ μF are connected to a battery of voltage 12 V as shown in Fig. 2.57. Find the charge on each capacitor when: (a) switch S_1 is closed, and (b) switch S_2 is also closed.

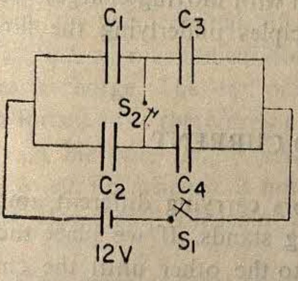


Fig. 2.57

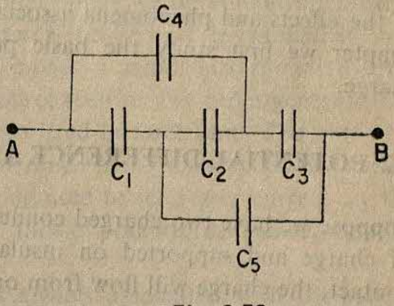


Fig. 2.58

25. In Fig. 2.58 the capacitors C_1, C_3, C_4 and C_5 have a capacitance of 4 μF each and the capacitor C_2 has a capacitance of 10 μF. Find the effective capacitance between points A and B.
26. A parallel-plate capacitor of plate area 100 cm^2 and plate separation 1.0 cm is connected across a 100 V battery. The battery is then disconnected and a dielectric slab of thickness 0.50 cm and dielectric constant 7.0 is inserted between the plates. Calculate (a) the capacitance C_0 (before the slab is inserted), (b) the free charge Q, (c) the electric field strength E_0 in the gap, (d) the electric field strength E in the dielectric, (e) the potential difference V between the plates, and (f) the capacitance C (when the slab is inserted).

3

CURRENT ELECTRICITY

3.1 INTRODUCTION

So far we have discussed the phenomena associated with charges at rest (or in static equilibrium). The branch of physics that deals with charges at rest is called *electrostatics*. By the middle of the eighteenth century, electrostatics was a well-established branch of physics. Machines had been invented which could produce large amounts of charge by friction. We will now discuss the motion or *dynamics* of charges. A moving charge constitutes an electrical current. The branch of physics which deals with charges in motion is called *current electricity*.

In this chapter we introduce the concepts of electrical *current* and electrical *resistance*. Chapters 4, 5, 7, 8 and 9 are devoted to the study of the effects and phenomena associated with moving charges. In this chapter we first study the basic principles underlying the flow of charge.

3.2 POTENTIAL DIFFERENCE AND CURRENT

Suppose we have two charged conductors carrying different amounts of charge and supported on insulating stands. If we place them in contact, the charge will flow from one to the other until the charges on them are equalized. The fact that charge flows can be demonstrated by any charge-detecting device, such as an electroscope or a galvanometer. What causes the charge to flow? What determines the direction of the flow of charge? Questions such as these can be answered in terms of the *electrical potential* of a charged body.

Just as the direction of flow of a liquid is independent of its total quantity (it depends only on the level difference) and the direction of flow of heat does not depend on the total quantity of heat (it depends on the temperature difference), the direction of the flow of charge does not depend on the quantity of charge of the two bodies; it depends on the *potential difference* between them. Charge flows from a body having a higher electrical potential to a body with a lower electrical potential.

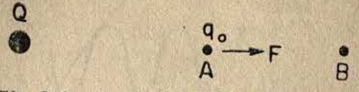


Fig 3.1 *Potential difference*

Consider two points A and B in the vicinity of a charge Q (Fig. 3.1). Let the potential at A be V_A and that at B be V_B. Since more work is needed to place a 'test' charge q_0 at A than at B, V_A is greater than V_B. If the charge placed at A is released, it will flow to B due to a potential difference ($V_A - V_B$). If A and B are two points on a conductor and a potential difference is maintained between them, then charge will flow from A to B. A cell serves this purpose.It provides the necessary energy to maintain a difference of potential. The cell was invented in 1791 by Luigi Galvani of Italy and greatly improved by Alessandro Volta in 1799. The working of a cell will be discussed in Chapter 4. We now know of many other devices that perform this function. A rectifier (described in Chapter 14) which converts the AC of the mains to DC is used in place of a cell.

Electromotive Force (EMF)

Any device used for moving a charge in a conductor has two terminals (positive and negative). The potential difference or voltage that exists between the two terminals of a source is called its *electromotive force* (emf). Just as mechanical force moves a mass, electromotive force moves a charge. The emf of a source (cell or any other generator) is not a force and is therefore not measured in force units. *The emf of a source is the potential difference between its terminals when it is on an open circuit*, i.e. when it is not being used to send any current in the circuit. Hence emf is measured in units of potential difference, namely, *volts*.

An electromotive force (emf) is always associated with the source (cell or any other generator) which makes the charge move from one place to another. Obviously, the source has to do work (or deliver energy) to move the charge. In an electrochemical cell, the energy released in a chemical reaction is transferred to the charges. Thus, as the reaction proceeds at a certain rate, the charges move from one electrode to another at a corresponding rate; each charge gaining a fixed potential energy in the process. In an electric generator, the source is the magnetic field changing with time. As discussed in Chapter 7, this produces an emf between two points (leads) of the generator.

We can define *the emf of a cell* (or a generator) *as the total energy per unit charge when the cell is on an open circuit, i.e. when the current through the cell is zero.* The SI unit of emf is *joules per coulomb*, which is the same as *volts*.

ELECTRIC CURRENT

When charge moves in a conductor, it does so at a certain rate which depends upon the emf of the source and the electrical property called *resistance* (see Sec. 3.3) of the conductor. The rate at which charge flows is called electric current. Quantitatively, electric current is defined as the amount of charge flowing through any cross-section of a conductor per second, i.e.

$$\text{Electric current} = \frac{\text{charge}}{\text{time}}$$

or

$$I = \frac{q}{t} \tag{3.1}$$

In the SI system, the unit of current is the *ampere*, denoted by symbol A. *The current is said to be one ampere when one coulomb of charge flows past any cross-section of the conductor every second.* When a current of two amperes flows through a conductor, two coulombs of charge pass through any cross-section of the conductor every second (see Fig. 3.2).

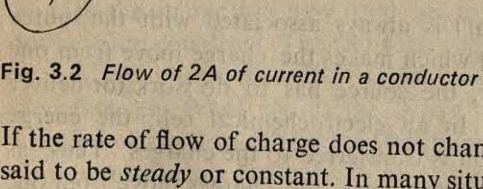

Two coulombs of charge pass through this cross-section (or any other cross-section every second)

Fig. 3.2 *Flow of 2A of current in a conductor*

If the rate of flow of charge does not change with time, the current is said to be *steady* or constant. In many situations, however, the current may vary with time. In such situations, the current is given by the differential limit of Eq. (3.1), or

$$I = \frac{dq}{dt} \tag{3.2}$$

In Fig. 3.3 the graph (a) represents a steady current while curves (b) and (c) represent varying currents. In this chapter we shall consider only steady currents.

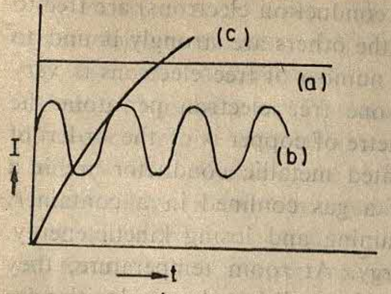


Fig. 3.3 *Steady and varying currents*

It may be mentioned that the current I is the same for all cross-sections of the conductor, even though the cross-sectional area may be different at different points. This may be understood from the analogy of water flowing through a pipe of non-uniform cross-sectional area. The water flows faster where the pipe is narrower and slower where the pipe is broader, but the rate of flow measured in litres per second remains unchanged. The constancy of current follows from the fact that charge must be conserved; it cannot pile up anywhere in the conductor.

Convention Regarding the Direction of Current

In metallic conductors the charge is carried by electrons. In electrolytes or gaseous conductors, the charge carriers may be positive or negative ions or both. A convention for labelling the direction of current is therefore needed. For simplicity and consistency, we assume the direction of current to be the one in which a positive charge would move. A positive charge moving in one direction is equivalent to a negative charge moving in the opposite direction. If the charge carriers are negative, they simply move opposite to the direction of the current arrow (see Fig. 3.4).

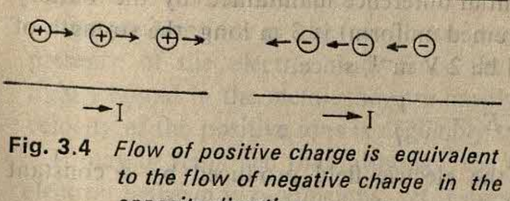


Fig. 3.4 *Flow of positive charge is equivalent to the flow of negative charge in the opposite direction*

3.3 MECHANISM OF FLOW OF CURRENT IN A METALLIC CONDUCTOR

We shall now examine the mechanism of the flow of current in a metallic conductor. In a metal in the solid state, such as a copper wire, the atoms are strongly bound to one another and occupy fixed positions. Some electrons (called the conduction electrons) are free to move in the body of the metal while the others are strongly bound to their atoms. In good conductors, the number of free electrons is very large. Assuming that there is only one free electron per atom, the number of free electrons per cubic metre of copper is of the order of 10^{28}. The free electrons in an isolated metallic conductor are in a random motion like the molecules of a gas confined in a container. They keep colliding with atoms, gaining and losing kinetic energy. Their motion is due to thermal energy. At room temperature, they move with velocities of the order of 10^5 m s^{-1}, but these velocities are randomly distributed in all directions. There is no net directed motion in any particular direction. If $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_3$, . . . , etc. are the velocities of individual electrons and if the total number of electrons is N, the average velocity of electrons is given by

$$\mathbf{u} = \frac{\mathbf{u}_1 + \mathbf{u}_2 + \mathbf{u}_3 + \ldots + \mathbf{u}_N}{N} = 0$$

In other words, there is no net flow of charge in any direction. If we consider any cross-section of the conductor, the rate at which electrons pass through it from right to left is the same as the rate at which they pass through it from left to right. The net rate is zero and no current flows in the conductor.

Drift Velocity

What causes current to flow in a metallic conductor? To answer this question, let us, for the sake of argument, assume that the conductor is in the form of a wire, say, of copper. If the ends of the wire are connected to a battery, an electric field will be set up at every point within the wire. If the potential difference maintained by the battery is 6 V and if the wire (assumed uniform) is 3 m long, the strength of this field at every point will be 2 V m^{-1}, since,

$$E = -\frac{V}{l}$$

where E is the strength of the electric field produced by a constant potential difference V applied across the wire of length l.

Under the influence of the electric field, the electrons experience coulomb force $-e\mathbf{E}$ in a direction opposite to that of the field $\mathbf{E}$ (since they have a negative charge). Each electron undergoes an acceleration $\mathbf{a}$ given by

$$\mathbf{a} = -\frac{e\mathbf{E}}{m}$$

where **E** is the electric field and e and m are respectively the charge and mass of an electron. These electrons which are loosely bound to their atoms break away and wander about in the body of the metal. As they get accelerated, they frequently collide with atoms (strictly ions) of the conductor. Between two successive collisions, an electron thus acquires (in addition to its thermal velocity) a velocity component in a direction opposite to **E**. The gain in velocity is $\mathbf{a}\tau$ where τ is the time interval between two successive collisions.

At any arbitrary instant of time, an electron which in the absence of the field **E**, has a thermal velocity $\mathbf{u}_1$ in any direction will now, in the presence of the field, have a velocity $(\mathbf{u}_1 + \mathbf{a}\tau_1)$ where τ_1 is the time that has elapsed since the last collision. Similarly the velocities of other electrons are $(\mathbf{u}_2 + \mathbf{a}\tau_2)$, $(\mathbf{u}_3 + \mathbf{a}\tau_3)$, . . . etc. The average velocity $\mathbf{v}_d$ of all the N electrons will be

$$\mathbf{v}_d = \frac{(\mathbf{u}_1 + \mathbf{a}\tau_1) + (\mathbf{u}_2 + \mathbf{a}\tau_2) + (\mathbf{u}_3 + \mathbf{a}\tau_3) + \ldots + (\mathbf{u}_N + \mathbf{a}\tau_N)}{N}$$

$$= \frac{\mathbf{u}_1 + \mathbf{u}_2 + \mathbf{u}_3 + \ldots + \mathbf{u}_N}{N} + \mathbf{a}\,\frac{\tau_1 + \tau_2 + \tau_3 + \ldots + \tau_N}{N}$$

$$= 0 + \mathbf{a}\tau$$

where $\tau = (\tau_1 + \tau_2 + \tau_3 + \ldots + \tau_N)/N$ represents the average time between two successive collisions. It is called the *relaxation time* and is of the order of 10^{-14} s for most conductors.

Thus $$\mathbf{v}_d = \mathbf{a}\tau = -\frac{e\mathbf{E}\tau}{m} \tag{3.3}$$

The parameter $\mathbf{v}_d$ is called the drift velocity of the electrons. It may be regarded as an average velocity of free electrons in a direction opposite to that of the field.

The positive ions in the conductor also experience a force in the presence of the electric field. Since the positive ions are heavy and tightly bound in the metal, they are hardly able to move and the drift velocity of the positive ions is negligibly small.

It may be remarked that although the electric field accelerates an electron between two collisions, yet it does not produce any net acceleration because the electron keeps colliding with ions of the metal. The electrons acquire a constant average drift velocity $\mathbf{v}_d$ in a direction opposite to **E** since the gain in velocity due to the field is lost in the next collision. The situation is similar to a small spherical metal ball rolling down a long flight of stairs. During the fall from one stair to the next, the ball acquires acceleration due to the gravitational force.

The moment it collides with the stair it is decelerated. The net result after a large number of steps is that the ball rolls down the stairs with zero average acceleration, that is, at constant average speed.

In conclusion, we may say that before the application of the electric field, the average velocity of electrons is zero and there is no net transport of charge in any direction. In other words, there is no flow of current in the conductor. After the application of the field, the electrons acquire a constant drift velocity in a direction opposite to **E** and there is a net transport of charge in a direction opposite to **E**. This is equivalent to saying that an electric current flows in the direction of **E** (See Fig. 3.5). It may be remarked that even in the case of good conductors, the magnitude of the drift velocity is not large and is usually about 1 mm s^{-1} (See Example 3.2.) But, since the number of electrons is very large (see Example 3.1), even this small drift of electrons adds up to a sizeable current.

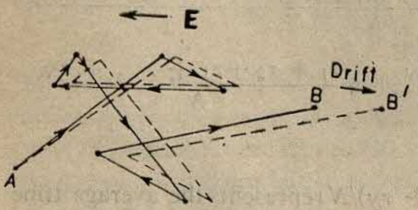


Fig. 3.5 *The solid lines show an electron moving from A to B, making six collisions. The broken lines show what the path of the electron might have been in the presence of an electric field* E. *Note the slow and steady drift in a direction opposite to* E.

Relation Between Current and Drift Velocity

Consider a conductor (e.g. a copper wire) of length l and of uniform cross-sectional area A. Let us say that a constant potential difference V is maintained across its ends with the help of a battery (Fig. 3.6).

The electric field E set up in the wire given by

$$E = -\frac{V}{l} \tag{3.4}$$

is set up in the wire. Under the influence of this field, the electrons drift with a constant velocity v_d in the direction opposite to that of E, i.e. towards the right as shown in the figure. In order to calculate the current I flowing in the wire, we need to compute the total charge

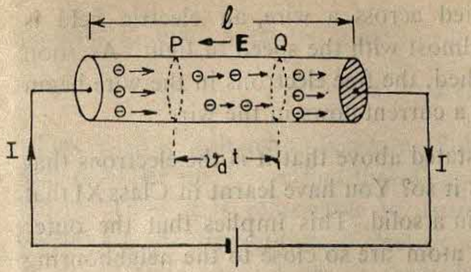


Fig 3.6 *The mechanism of flow of current in a conductor*

passing through any cross-section of the wire in a certain interval of time, say, t. In time t, each electron moves through a distance $v_d t$ towards the right. If we consider two planes P and Q separated by a distance $v_d t$, then the total charge flowing in time t is the charge of the electrons contained in a cylinder PQ. If n is the number of electrons per unit volume, then the number of electrons contained in cylinder PQ $= n$ times the volume of the cylinder PQ $= n \times A \times v_d t$. If the charge on the electron is e, the total charge q flowing through any cross-section of the wire in time t is given by

$$q = enAv_d t$$

Hence the current I in the wire is given by

$$I = \frac{q}{t} = enAv_d \tag{3.5}$$

As mentioned earlier, the magnitude of the drift velocity in metallic conductors is very small (~ 1 mm s^{-1}). With this speed, the electrons will take about twenty minutes to drift through a wire of length 1 m. It may appear to be surprising and often confusing that, although the drift velocity of electrons is so small, yet an electric bulb lights up almost instantly when it is switched on. The fact is that an electrical impulse propagates with a very high speed which is of the order of the speed of light ($= 3 \times 10^8$ m s^{-1}). Our analogy of water in a pipe again comes to our rescue and helps clear the confusion. When pressure is applied by pushing a piston at one end of a long water filled pipe, we observe that water starts coming out of the other end almost immediately. As soon as the pressure is applied, a compression wave is transmitted rapidly along the pipe. The speed of this wave is about 1400 m s^{-1}. When this wave reaches the other end, the flow of water starts. The water inside the pipe also starts moving forward, but this speed is much slower than that of the pressure wave. Similarly, when a potential difference is applied across a wire, an electric field is established in the whole wire almost with the speed of light. As soon as the electric field is established, the free electrons in the wire begin to drift under its influence and a current flows in the wire.

Carriers of Current We have stated above that it is the electrons that carry current in metals. Why is it so? You have learnt in Class XI that atoms are very closely packed in a solid. This implies that the outer (or valence) electrons of one atom are so close to the neighbouring atoms that they do not stay 'tied' to a particular atom and can, therefore, roam around 'freely' in the entire body of the solid. These 'free' electrons carry electrical current.

This simple picture is, however, not a complete picture. It does not, for instance, explain why some solids are metals (i.e. conductors) while others are insulators. The answer to this question can be given in terms of the wave nature of electrons, as we will learn in Chapter 12. The simple rule is this: if the number of valence electrons per unit cell of the crystalline solid is *odd*, the solid is a *metal*; but if this number is *even*, the solid *can* be an *insulator* or even a *metal*.

Consider the elements in the first group (Li, Na, K, Rb, Cs, Cu, Ag, Au). They all have one atom in a unit cell and one valence electron per atom. Thus, according to the above rule, these solids should be metals, as they indeed are. Now consider the elements in the fourth group (C, Si, Ge, Sn, Pb). They all have one atom in a unit cell and four valence electrons per atom. Our rule tells us that these solids with an even number of valence electrons per unit cell can be insulating or metallic. Thus diamond (a form of carbon) is an insulator and Si and Ge are semiconductors but Sn and Pb are conductors (metals).

Distribution of Electron Velocities

The free electrons in metals move about with different speeds in different directions. If $N(v)$ denotes the number of free electrons with speed v, then the distribution $N(v)$ of electrons as a function of speed v for a typical metal is as shown in Fig. 3.7. The number of electrons with speed between v and $v + \Delta v$ is proportional to $N(v)\Delta v$.

The maximum speed v_0 of electrons in a typical metal is of the order of 10^6 ms^{-1}. The corres-

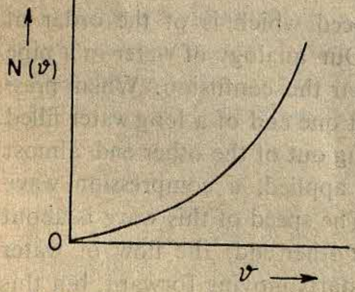


Fig. 3.7 *Variation of N(v) with v for free electrons in a metal*

ponding kinetic energy is related to the absolute temperature T_0 as

$$\frac{1}{2}mv_0^2 = kT_0$$

where k is the Boltzmann constant. Substituting the values of m, k and $v_0 = 10^6\ \mathrm{ms^{-1}}$, we find that T_0 is of the order of a few thousand kelvin. Thus it is reasonable to assume that at room temperatures (~300 K), the distribution of electron speeds is very nearly the same as at absolute zero. Even at these speeds the free electrons do not 'pop' out of the metal because they are attracted to the ions in the solid and also because of surface forces. Energy has to be supplied to pull an electron completely out of a metal. This energy is called the *work function* of the metal (see Ch. 12).

EXAMPLE 3.1 A current of 1 A flows through a copper wire. Calculate the number of electrons passing through any cross-section of the wire in 1.6 s? Given the charge on an electron = 1.6×10^{-19} C.

Solution: Since the current is 1 A, the charge passing through any cross-section of the wire in 1 s is 1 C. In 1.6 s, 1.6 C of charge is transferred. Now

$$q = Ne$$

where $q = 1.6\mathrm{C}$, $e = 1.6 \times 10^{-19}$ C and N = number of electrons

$$\therefore \qquad N = \frac{q}{e} = \frac{1.6\mathrm{C}}{1.6 \times 10^{-19}\mathrm{C}} = 10^{19}$$

This is an enormous number. In copper there is one free electron per atom, the other 28 electrons of a copper atom remain bound to it. A copper wire is made up of a very large number of atoms and hence has a very large number of free (conduction) electrons.

EXAMPLE 3.2 A uniform copper wire of diameter 0.050 cm carries a current of 1.0A. (a) Calculate the average drift speed of conduction electrons in the wire, assuming that there are 8.4×10^{28} free electrons/$\mathrm{m^3}$ in copper. (b) Compare the drift speed with (i) the thermal speed of copper atoms at room temperature of 27°C and (ii) the maximum speed of electrons carrying the current. Given, charge on an electron $(e) = 1.6\ 10^{-19}$ C, mass of an electron $(m) = 9.1 \times 10^{-31}$ kg, Boltzmann constant $(k) = 1.38 \times 10^{-23}\ \mathrm{JK^{-1}}$.

Solution: (a) We know that the average drift speed of the conduction electrons is given by

$$v_\mathrm{d} = \frac{I}{eNA} \qquad \text{(i)}$$

Now $I = 1.0$ A, $n = 8.4 \times 10^{28}$ electrons/m^3 and $A = \pi r^2 = 3.14 \times (0.025 \text{ cm})^2 = 1.9 \times 10^{-3} \text{ cm}^2 = 1.9 \times 10^{-7} \text{ m}^2$. Substituting the values of I, n, A and e in (i) gives

$$v_d = 3.9 \times 10^{-4} \text{ ms}^{-1}$$

(b) (i) The average thermal speed of copper atoms at absolute temperature T is given by

$$v_t = \sqrt{\frac{kT}{m}} \qquad \text{(ii)}$$

Substituting the values of m, k and $T = 273 + 27 = 300$ K in (ii), we get

$$v_t = 6.7 \times 10^4 \text{ ms}^{-1}$$

$$\therefore \qquad \frac{v_d}{v_t} = \frac{3.9 \times 10^{-4}}{6.7 \times 10^4} \simeq 5.8 \times 10^{-7}$$

Thus the drift speed of electrons is about 10^{-7} times smaller than the typical thermal speed at ordinary temperatures.

(ii) The maximum speed (v_0) of electrons corresponds to a temperature $T_0 \simeq 10^5$ K. Thus

$$v_0 = \sqrt{\frac{kT_0}{m}} \qquad \text{(iii)}$$

Substituting the values of k, m and T_0 in (iii) we get

$$v_0 = 1.4 \times 10^6 \text{ ms}^{-1}$$

which is about 10^9 times smaller than the drift speed.

3.4 OHM'S LAW: RESISTANCE AND RESISTIVITY

In an electrical circuit, the cell is the source of electrical energy which maintains the desired potential difference between the ends of a conductor. It is the potential difference that drives the current in the circuit. We would expect a definite relation between the potential difference between the ends of a conductor and the magnitude of the current that flows in it. This very important relation was discovered by George Simon Ohm (Germany) in the year 1827. On the basis of his experimental observations he discovered a law, know in his honour as *Ohm's Law*, which states that *the current flowing through a conductor is directly proportional to the potential difference across its ends, provided the physical conditions of the conductor remain the same.*

We shall later discuss what physical conditions of the conductor must remain the same for Ohm's law to hold. The above relationship tells us that if the potential difference across the ends of a conductor is doubled, the current flowing in it will also be doubled. Thus,

$$\text{potential difference} \propto \text{current}$$

or $$V \propto I$$

or $$V = RI \tag{3.6}$$

where R is the constant of proportionality called the resistance of the given conductor.

Unit of Resistance The SI unit of electrical resistance is called the *Ohm* (denoted by symbol Ω). If the potential difference (V) is 1 volt and current (I) is 1 ampere, the resistance (R) is 1 ohm. From Ohm's law we have

$$R = \frac{V}{I}$$

$$\therefore \qquad 1 \text{ ohm} = \frac{1 \text{ volt}}{1 \text{ ampere}}$$

Thus, *the resistance of a conductor is said to be 1 ohm if a current of 1 ampere flows through it when the potential difference across its ends is 1 volt.*

What Causes Resistance in a Conductor

Exactly how much current will flow through a conductor, for a given potential difference across its ends, depends upon how much resistance the conductor offers to the flow of electrons. Let us suppose that the conductor is in the form of a wire. Electrons frequently collide with the atoms of the wire as they move through it. These collisions slow down the electrons. The resistance of a wire depends upon the number of collisions the electrons make with the atoms of the wire. Therefore resistance depends on the arrangement of atoms in the material, i.e., on the kind of material (copper, silver etc.). Resistance must also depend upon the length and thickness of the wire. A long wire will offer more resistance than a short one because there will be more collisions in the longer wire. Also a thick wire will offer less resistance than a thin wire because in the thick wire there is more area of cross-section for the electrons to flow through. Thus we conclude that the resistance of a wire depends upon: (a) the material of the wire, (b) its length, and (c) its thickness.

Electrical Resistivity of Materials

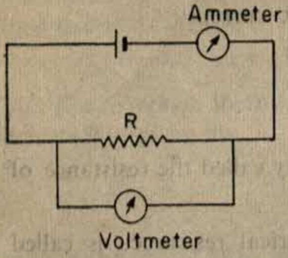


Fig. 3.8 *Measurement of resistance of a wire*

The resistance of a wire can be measured by a simple experiment illustrated in Fig. 3.8. The ratio of the voltmeter reading (V) and the ammeter reading (I) determines the resistance (R) of the wire.

The experiment shows that the resistance of two wires of different material, say copper and silver, having the same length and the same diameter (i.e. the same cross-sectional area) is different. This shows that resistance depends on the material of which the wire is made. The experiment also shows that if we take two wires of the same material and having the same diameter, their resistances will be different if they have different lengths. In fact, if the length is doubled, the resistance is also doubled, i.e. resistance R is directly proportional to the length of the wire.

$$R \propto l \tag{3.7}$$

The experiment also reveals that if we take two wires of the same material and having the same length, their resistance will be different if they have different areas of cross-section. In fact, if cross-sectional area A is doubled, the resistance R becomes one-half, i.e. resistance is inversely proportional to the area of the cross-section of the wire.

$$R \propto \frac{1}{A} \tag{3.8}$$

Combining (3.7) and (3.8) we have

$$R \propto \frac{l}{A} \quad \text{or} \quad R = \rho \frac{l}{A}$$

or

$$\rho = \frac{RA}{l} \tag{3.9}$$

where ρ is a constant of proportionality and its value depends on the material of the wire. The constant ρ is called the *resistivity* of the material of the wire. From Equation (3.9) we find that if,

$$l = 1 \text{ m}$$

and

$$A = 1 \text{ m}^2$$

then

$$\rho = R$$

Thus the *resistivity of the material of a wire is the resistance of a wire of length 1 m and area of cross-section equal to 1 m²*.

Unit of ρ : From Eq. (3.9) we have

$$\text{Unit of } \rho = \frac{\text{Unit of } R \times \text{unit of } A}{\text{unit of } l}$$

$$= \frac{\text{ohm} \times (\text{metre})^2}{\text{metre}}$$

$$= \text{ohm metre}$$

Hence ρ is measured in ohm metre (or Ω m).

Note that the value of ρ is independent of the length and the area of cross-section of the wire; it depends only on the material of the wire. Thus, resistivity is a characteristic of the material. Hence we can compare resistances of various substances if we know their resistivity. Table 3.1 shows resistivities of various substances.

Table 3.1 Electrical resistivities of some substances

Material	Resistivity (ρ) at 0°C (Ωm)	Temperature coefficient $\alpha = \frac{1}{\rho}\left(\frac{d\rho}{dT}\right)$ (k^{-1})	Number of valence electrons per unit cell
A. *Conductors*			
Copper	1.7×10^{-8}	3.9×10^{-3}	1
Silver	1.6×10^{-8}	4.1×10^{-3}	1
Aluminium	2.7×10^{-8}	4.3×10^{-3}	3
Tungsten	5.6×10^{-8}	4.5×10^{-3}	6
Iron	1.0×10^{-7}	6.5×10^{-3}	8
Platinum	1.1×10^{-7}	3.9×10^{-3}	10
Nichrome	1.0×10^{-6}	9×10^{-4}	—
B. *Semiconductors*			
Carbon	3.5×10^{-5}	-5×10^{-4}	4
Germanium	0.46	-5×10^{-2}	4
Silicon	2300	-7×10^{-2}	4
C. *Insulators*			
Glass	$10^{10} - 10^{14}$	–	–
Hard rubber	$10^{13} - 10^{16}$	–	–

The reciprocal of electrical resistivity ($1/\rho$) is called *electrical conductivity* (σ), i.e. $\sigma = 1/\rho$; σ is expressed in ohm^{-1} m^{-1} or mho m^{-1}. Table 3.1 shows that the electrical resistivity of substances varies over a very wide range. Metals have low resistivities (and hence high conductivities). On the other hand substances such as glass and hard rubber have very high resistivity (hence negligibly low conductivity); they are insulators. In between lie semiconductors, such as silicon and germanium.

Derivation of Ohm's Law for Conductors: Nature of Electrical Resistance

We can understand and deduce Ohm's law as follows:

We have seen that the average drift speed $\mathbf{v}_d$ of conduction electrons depends on the strength of the electric field **E** [Eq. (3.3)] which in turn depends on the potential difference V as is clear from Eq. (3.4). In terms of V, the magnitude v_d of the drift velocity can be obtained by using Eq. (3.4) in Eq. (3.3)

$$v_d = \frac{eV\tau}{ml} \tag{3.10}$$

Substituting for v_d from Eq. (3.10) in Eq. (3.5) we get

$$I = (enA)\left(\frac{eV\tau}{ml}\right)$$

$$= \left(\frac{ne^2A\tau}{ml}\right)V$$

or

$$I = \frac{V}{R} \tag{3.11}$$

where $R = ml/(ne^2A\tau)$ is a constant for a given conductor.

Equation (3.11) is a statement of Ohm's law for conductors. The constant R is called the resistance of the conductor and its value depends on the dimensions (A and l) of the wire and on the material (n and τ) of the wire. We can also write R as

$$R = \rho\frac{l}{A} \tag{3.12}$$

where the electrical resistivity ρ of the material is given by

$$\rho = \frac{m}{ne^2\tau} \tag{3.13}$$

Notice that ρ is independent of the dimensions of the conductor and depends only on its material.

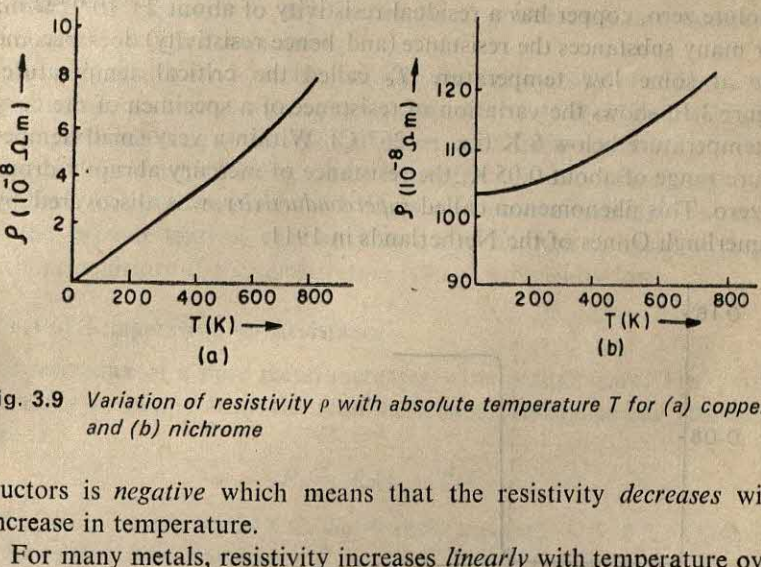
Variation of Resistivity with Temperature

Equation (3.13) shows how the resistivity ρ is related to two parameters of the material, namely the number density n of the free electrons and mean collision time τ. The former obviously cannot change with temperature but the latter does. The thermal speed of electrons increases with increase in temperature. Consequently the collisions become more frequent, resulting in a decrease in collision time τ. Hence *resistivity* [see Eq. (3.13)] *increases with increase in temperature.*

The temperature coefficient α of resistivity is given by

$$\propto = \frac{1}{\rho}\frac{d\rho}{dT} \tag{3.14}$$

and is *defined as the fractional change in resistivity* ($d\rho/\rho$) *per unit change in temperature* (dT). The value of α also varies with temperature. In Table 3.1, the values of α refer to 20°C. For conductors, α is *positive* indicating that α *increases* with temperature. Note that α for semiconductors is *negative* which means that the resistivity *decreases* with increase in temperature.



(a)

(b)

Fig. 3.9 *Variation of resistivity ρ with absolute temperature T for (a) copper and (b) nichrome*

For many metals, resistivity increases *linearly* with temperature over a temperature range of about 500°C above the room temperature (see Fig. 3.9). Over this range of temperatures, the value of α for many metals (such as copper) is a constant, independent of temperature; depending only on the nature of the metal.

In insulators, the resistivity increases exponentially with decrease in temperature, becoming infinitely large at temperatures near absolute zero. Such substances whose resistance becomes infinite at $T = 0$ are called insulators. Some semiconductors also show this behaviour. The temperature dependence of resistivity of insulators is given by

$$\rho = \rho_0 \exp\left(\frac{E_g}{kT}\right)$$

where k is the Boltzmann constant and E_g is a positive energy. The large (exponential) variation of resistivity is due to the fact that E_g can vary from nearly zero value to several electron volts. At room temperature ($T \sim 300$ K), $kT \simeq 0.03$ eV. Hence E_g/kT can vary from a very small value ($\sim$ zero) to about a 100 or more. It is the value of the exponent E_g/kT which distinguishes an insulator from a semiconductor. If $E_g \leqslant 1$ eV, the substance is a semiconductor at room temperature, then the resistivity at room temperature is not very high. But if $E_g \geqslant 1$ eV, the resistivity at room temperature is very high ($\sim 10^3\ \Omega$ m) and the substance is an insulator.

Superconductivity The curve of Fig. 3.9 (a) does not really go to zero at absolute zero of temperature (although it appears to do so). At absolute zero, copper has a residual resistivity of about $2 \times 10^{-10}\ \Omega$ m. For many substances the resistance (and hence resistivity) does become *zero* at some low temperature T_c called the critical temperature. Figure 3.10 shows the variation of resistance of a specimen of mercury at temperature below 6 K (i.e. $-$ 267°C). Within a very small temperature range of about 0.05 K, the resistance of mercury abruptly drops to zero. This phenomenon called *superconductivity* was discovered by Kamerlingh Onnes of the Netherlands in 1911.

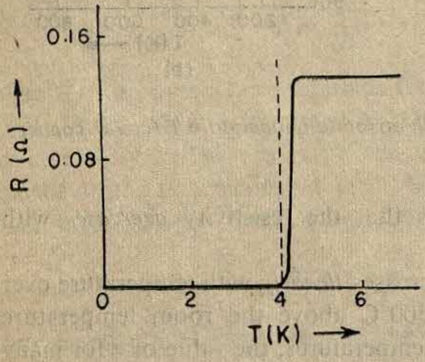


Fig. 3.10 *The resistance of mercury disappears below* 4K

Mercury metal becomes a superconductor at 4 K (liquid helium temperature), i.e. its resistance entirely disappears at a temperature of 4 K, which is called the *superconducting point*. The resistance of metallic tin falls to zero near 3.72 K and that of Nb_3Sn near 18.2 K. Many elements and several hundred compounds show this kind of behaviour.

Superconductors are the most exotic of all conductors. The fact that the resistance of a substance becomes zero at a certain low temperature has aroused the curiosity of both experimental and theoretical physicists. The zero resistance of a semiconductor implies that a current flows with no expenditure of energy. Once a current is set up in a loop of a semiconducting material, it persists unchanged for weeks and even months, even though there is no source of emf (battery) in the circuit. This remarkable property of superconductors will no doubt have very important practical applications. A serious difficulty in their use is the very low temperatures at which they must be kept. Recent newspaper reports (1987–88) say that scientists all over the world are trying to achieve superconductivity at room temperatures. If and when this is achieved, we would be able to transmit electrical power over superconducting cables without any loss of power during transit. We would also be able to produce and maintain very high magnetic fields without the expenditure of megawatts of power as is now necessary in nuclear particle accelerators.

It is not clearly understood yet why, under certain conditions, the conduction electrons can move completely freely through matter. In the superconducting state, the electrons are acted upon by a weak attractive force which brings them together to form coupled pairs. On the basis of quantum mechanical ideas (which are outside the scope of the present text) it can be shown that the paired electrons move without collisions if the temperature is kept sufficiently low.

Effect of Temperature on Resistance

The resistance of a pure metal increases with temperature. For small increases in temperature the relation between resistance and temperature is

$$R_t = R_0(1 + \beta t)$$

where R_t = resistance at t°C, R_0 = resistance at 0°C and t = rise in temperature. The parameter β is called the temperature coefficient of resistance. For pure metals, β is positive. For copper, for instance, $\beta = 0.0038$ per degree Celsius. For alloys, such as manganin, eureka, etc., β is very small $\simeq 0.00001/°\mathrm{C}$ implying that the change in resistance with temperature is negligible for a fairly large temperature difference. It is for this reason that these alloys are used to construct standard resistances. For carbon and some semiconductors, β is negative, indicating that their resistance decreases with increase in temperature. Ohm's law does not apply in such cases.

Validity of Ohm's Law: Ohmic and Non-ohmic Conductors

It must be clearly understood that ohm's law is not a law of nature.

The law is obeyed by metallic conductors; they are, therefore, called ohmic conductors. If we apply a variable potential difference across a metal wire, then for each applied potential difference V, *the current I in the wire is proportional to the applied potential difference and the resistance of the wire is the same, no matter what the applied voltage is.* If V is reversed, I is also reversed but the magnitude of I remains unchanged. The characteristic I-V graph for metals is a straight line passing through the origin (Fig. 3.11). We assume that the temperature of the conductor is kept constant. But at higher currents, the conductor becomes hot and its resistance increases. Thus the I-V graph is linear only for small currents, as shown in Fig. 3.11 where the dashed line represents linear behaviour and the solid curve at higher currents shows a departure from linear behaviour.

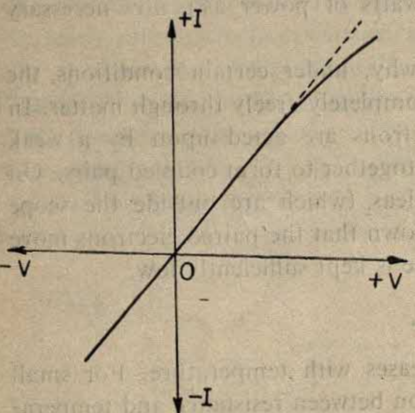


Fig. 3.11 *Characteristic I-V graph for ohmic conductors*

Many conductors do not obey Ohm's law; they are called non-ohmic conductors. The I-V graph of such conductors has one or more of the following properties:

(a) The relation between I and V is nonlinear.

(b) The relation between I and V depends upon the sign of V for the same absolute value of V.

(c) The relation between I and V is not unique, i.e. I may have more than one value at a given value of V.

Many useful components in electrical circuits are required to be non-ohmic. For example, a non-ohmic component is essential in a radio receiver circuit. Figure 3.12(a) shows the I-V characteristic curve for a junction (semiconductor) diode used as a rectifier; it exhibits the properties (a) and (b) listed above. Notice that I is not proportional to V. Further, very little current flows even for fairly high negative voltage V (called negative bias) and a current begins to flow for much smaller positive bias. Thus the diode junction allows current to flow in only one direction, i.e. it acts as a *rectifier*. In Chapter 14, we will learn more details of a rectifier which is a very useful and essential circuit element.

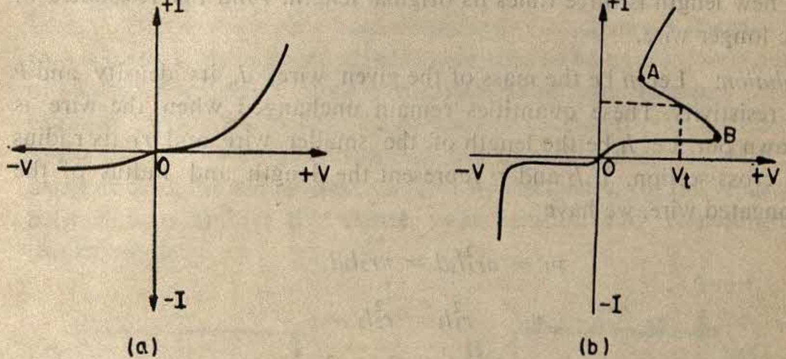


Fig. 3.12 *I-V characteristics of (a) a diode junction and (b) a thyrister*

Figure 3.12(b) shows the characteristics of a circuit element known as the *thyrister* which consists of four alternate layers of *p*- and *n*-type semiconductors; it exhibits all the three properties (a), (b) and (c) of non-ohmic conductors. Notice that again I is not proportional to V. Further, the current at a given voltage is not unique; it has two values at a voltage V_t. The region AB is interesting because over this region the current increases with decrease in voltage, completely violating Ohm's law.

EXAMPLE 3.3 In a discharge tube, the number of protons drifting across any cross-section is 1.0×10^{18} per second while the number of electrons drifting in the opposite direction across another cross-section is 2.7×10^{18} per second. If the voltage across the tube is 220 V, what is the effective resistance of the tube?

Solution: Let us say that the protons are drifting in the $+x$ direction. The current due to the drifting of protons is $(1.0 \times 10^{18} \times 1.6 \times 10^{-19})$ Cs^{-1} (or A) in the $+x$ direction. Since the electrons carry a negative charge and they are drifting in the $-x$ direction, the current due to electrons is $2.7 \times 10^{18} \times (-1.6 \times 10^{-19})$ A in the $-x$ direction which is equivalent to a current of $(2.7 \times 10^{18} \times 1.6 \times 10^{-19})$ A in the $+x$ direction. Hence the effective current in the $+x$ direction is

$$I = 3.7 \times 10^{18} \times 1.6 \times 10^{-19} = 5.92 \times 10^{-1} \text{ A}$$

Now $V = 220$ V. The effective resistance of the tube is

$$R = \frac{V}{I} = \frac{220 \text{ V}}{5.92 \times 10^{-1} \text{ A}}$$

$$= 3.7 \times 10^{2} \Omega$$

EXAMPLE 3.4 A wire with a resistance of 5 ohms is drawn out so that its new length is three times its original length. Find the resistance of the longer wire.

Solution: Let m be the mass of the given wire, d, its density and ρ its resistivity. These quantities remain unchanged when the wire is drawn out. Let l_1 be the length of the smaller wire and r_1 its radius of cross-section. If l_2 and r_2 represent the length and radius of the elongated wire, we have

$$m = \pi r_1^2 l_1 d = \pi r_2^2 l_2 d$$

i.e. $$r_1^2 l_1 = r_2^2 l_2$$

It is given that $l_2 = 3l_1$. Therefore, $r_2^2 = r_1^2/3$.

If R_1 is the resistance of the smaller wire and R_2 that of the elongated wire, we have

$$R_1 = \frac{\rho l_1}{\pi r_1^2}$$

and $$R_2 = \frac{\rho l_2}{\pi r_2^2}$$

so that $$\frac{R_2}{R_1} = \frac{l_2}{l_1} \cdot \frac{r_1^2}{r^2} = 9$$

Since $$R_1 = 5 \, \Omega$$

$$R_2 = 45 \, \Omega$$

EXAMPLE 3.5 A copper coil has a resistance of 2.00 Ω at 20°C. Find its resistance at 50°C. The temperature coefficient of resistance for copper is 3.80×10^{-4} per degree Celsius.

Solution: We know that

$$R_t = R_0 (1 + \beta t)$$

$$\therefore \quad R_{20} = R_0 (1 + 20 \, \beta)$$

$$R_{50} = R_0 (1 + 50 \, \beta)$$

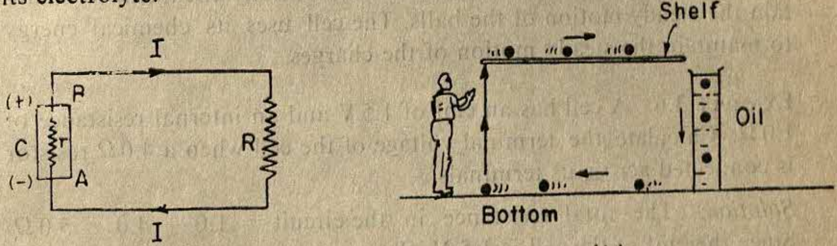
$$\therefore \quad \frac{R_{50}}{R_{20}} = \frac{1 + 50\,\beta}{1 + 20\,\beta}$$

Now $R_{20} = 2.00\,\Omega$ and $\beta = 3.80 \times 10^{-4}$ per °C. Hence

$$R_{50} = 2.02\,\Omega$$

3.5 INTERNAL RESISTANCE

Let us consider a simple circuit consisting of a resistance R connected to a source of emf labelled C (which we take to be a cell). Every source of emf has a resistance of its own called the *internal resistance* (r). For example, in the case of cell, the components inside it offer a resistance to the flow of current. The value of the internal resistance of a cell depends on: (a) the surface area of its electrodes, (b) the separation between them and (c) the nature, concentration and temperature of its electrolyte.



(a) (b)

Fig. 3.13 *(a) A simple electric circuit and (b) its gravitational analogue*

When a current I flows through C, the source C does work to transfer positive charge from one end A to the other end B. If $\mathcal{E}$ is the emf of the cell and r its internal resistance, then a part v of $\mathcal{E}$ is used up to transfer charge from A to B; from Ohm's law this part is given by

$$v = Ir \tag{3.14}$$

The remaining part $V = (\mathcal{E} - v)$ drives the charge through the external resistor R; this part is given by

$$V = IR \tag{3.15}$$

Now

$$\mathcal{E} = V + v \tag{3.16}$$

Equations (3.14) to (3.16) give

$$V = \frac{\mathcal{E}r}{(R + r)}$$

and

$$I = \frac{\mathcal{E} - V}{r} \tag{3.17}$$

Notice from Eq. (3.17) that if $I = 0$, $\mathcal{E} = V$, i.e. if no current is drawn from the cell (i.e. it is on an open circuit), the potential difference V across its terminals equals its emf $\mathcal{E}$. If $I \neq 0$, V is less than $\mathcal{E}$.

Figure 3.13(b) shows a gravitational analogue of Fig. 3.13(a) which helps one understand the concept of emf. The man lifts balls from the bottom and transfers them to the top of a shelf by doing work on them; this energy is stored temporarily as gravitational energy. In Fig. 3.13(a) the cell transfers charge from terminal A to B at the expense of its chemical energy which is stored temporarily as electrical potential energy. The balls roll slowly and uniformly along the shelf and drop from the right end into a cylinder filled with viscous oil. They sink to the bottom at a constant speed. In Fig. 3.13(a) the charges move through the resistance R at a constant speed. The balls roll back along the floor. The balls move through the system at a steady rate just like charge moving through a circuit, making a steady current. The man uses his own energy (which is essentially chemical) to maintain the steady motion of the balls. The cell uses its chemical energy to maintain the steady motion of the charges.

EXAMPLE 3.6 A cell has an emf of 1.5 V and an internal resistance of 1.0 Ω. Calculate the terminal voltage of the cell when a 4.0 Ω resistor is connected across its terminals.

Solution: The total resistance in the circuit = 1.0 + 4.0 = 5.0 Ω. Since the emf of the cell is 1.5 V, the current in the circuit is

$$I = \frac{1.5\ \text{V}}{5.0\ \Omega} = 0.3\ \text{A}$$

Terminal voltage (V) of cell = potential difference across the 4.0 Ω resistor

$$= 0.3 \times 4.0 = 1.2\ \text{V}$$

EXAMPLE 3.7 A battery supplies a current of 0.9 A through a 2.0 Ω resistor and a current of 0.3 A through a 7.0 Ω resistor. Calculate the emf and the internal resistance of the battery.

Solution: Let $\mathcal{E}$ be the emf and r the internal resistance of the battery. Since emf = current × total resistance, we have

In the first case : $\mathcal{E} = 0.9(2.0 + r)$ (i)

In the second case: $\mathcal{E} = 0.3(7.0 + r)$ (ii)

Equating (i) and (ii) and solving for $\mathcal{E}$ and r we get

$$\mathcal{E} = 2.25\ \text{V} \qquad \text{and} \qquad r = 0.5\ \Omega$$

3.6 CELLS IN SERIES AND IN PARALLEL

The current obtained from a single cell is small. In experimental work, higher currents are often required. For this purpose a combination of two or more cells is used. The combination of cells is called a battery. The battery of cells is used in torches, transistor sets, cars and other automobiles, trains, etc. Cells can be combined in various ways. We shall first consider the simplest combination, namely, the series combination.

Cells are said to be combined in series if the positive terminal of one cell is joined to the negative terminal of the second and so on, as shown in Fig. 3.14. Let us consider n cells connected in series, each of emf $\mathcal{E}$ and internal resistance r. Let R be the external resistance.

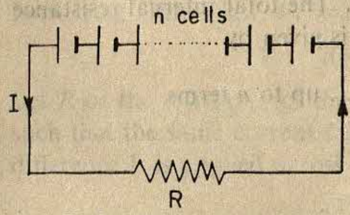


Fig. 3.14 *Cells connected in series*

Then, total emf of the battery $= n \times \mathcal{E}$ and total internal resistance of the battery $= n \times r$. Therefore,

$$\text{Total resistance in the circuit} = nr + R$$

Now, from Ohm's law

$$\text{Current } (I) \text{ in the circuit} = \frac{\text{total potential difference}}{\text{total resistance}}$$

or
$$I = \frac{n\mathcal{E}}{nr + R} \tag{3.18}$$

Let us consider two special cases:

1. $R \gg nr$. If the external resistance R is very large compared to the total internal resistance nr, then nr can be neglected compared to R and we have

$$I = \frac{n\mathcal{E}}{R} \tag{3.19}$$

which is n times the current $\mathcal{E}/R$ that can be drawn from a single cell.

Hence n cells combined in series multiply the current in a circuit by n times that given by a single cell if $R \gg nr$.

2. $R \ll nr$. On the other hand if $R \ll nr$ then we have

$$I = \frac{n\mathcal{E}}{nr} = \frac{\mathcal{E}}{r}$$

which is the same as that given by a single cell.

Thus we conclude that a series combination of cells should be used

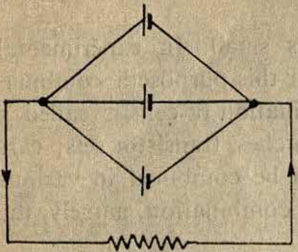


Fig. 3.15 *Cells connected in parallel*

if the external resistance is high compared to the total internal resistance of the battery. If this is not the case, then a parallel combination of cells is employed where the positive terminals of the cells are connected together and so are all their negative terminals, as shown in Fig. 3.15.

If identical cells are connected in parallel, the effective emf is equal to the emf of any one of the cells. The total internal resistance of n identical cells connected in parallel is given by

$$\frac{1}{r_n} = \frac{1}{r} + \frac{1}{r} + \ldots \text{ up to } n \text{ terms}$$

$$= \frac{n}{r}$$

or
$$r_n = \frac{r}{n} \tag{3.20}$$

where r is the internal resistance of each cell. The calculation of the total emf of a parallel combination of different cells is more complicated. This problem can be tackled by using Kirchhoff's rules described in Sec. 3.8.

3.7 RESISTANCES IN SERIES AND IN PARALLEL

In electrical circuits, it is often necessary to combine two or more resistances. Resistances can be connected in two ways: (a) in series, and (b) in parallel. We shall find out the resistance of the combination in each case.

1. Resistances in Series

Resistances are said to be connected in series when they are joined end-to-end. Figure 3.16 shows three resistances R_1, R_2 and R_3 connected in series. When a battery is connected as shown, a current flows in the circuit, the current is the same in all the resistances, since any charge flowing into the first resistance must next flow through the second, and so on.

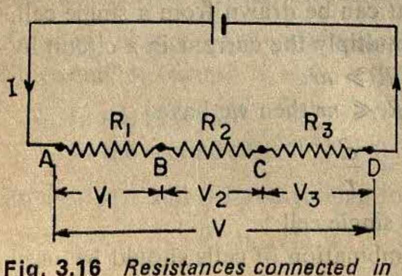


Fig. 3.16 *Resistances connected in series*

There is no other alternative path for the charge to flow. Thus *in a series circuit, the current is the same at all points.* Let V be the voltage of the battery. Therefore, the potential difference between the points A and D in the circuit is V volts. Let I be the current flowing in the circuit. Since the current flowing through each resistance is the same $(= I)$ and since the resistances R_1, R_2 and R_3 are different, Ohm's law $(V = IR)$ tells us that the potential difference across the ends of each resistance will be different.

If V_1, V_2 and V_3 are the potential differences across R_1, R_2 and R_3 respectively, then

$$V = V_1 + V_2 + V_3$$

Now from Ohm's law, $V_1 = IR_1$, $V_2 = IR_2$ and $V_3 = IR_3$. Therefore,

$$V = I(R_1 + R_2 + R_3) \tag{3.21}$$

Let R be the resistance of the combination. The value of R must be such that the same current I flows through it when the same potential difference V is applied across it, i.e.

$$V = IR \tag{3.22}$$

Equating (3.21) and (3.22), we get

$$R = R_1 + R_2 + R_3$$

For n resistances in series, we have

$$R = R_1 + R_2 + R_3 + \ldots + R_n \tag{3.23}$$

Thus *when resistances are connected in series, the total resistance is equal to the sum of the individual resistances.*

2. Resistances in Parallel

A circuit diagram showing three resistances R_1, R_2 and R_3 connected in parallel is shown in Fig. 3.17. The current I leaving the battery at its positive terminal divides at the junction into three parts I_1, I_2 and I_3 which recombine at other junction to give finally the same total current returning to the battery. It is clear that,

$$I = I_1 + I_2 + I_3$$

It is evident from Fig. 3.17 that the potential difference across each resistance is V, the voltage of

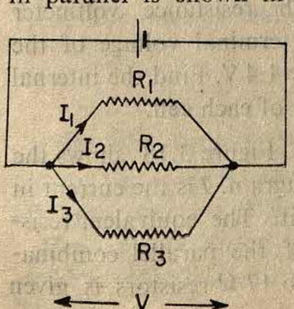


Fig. 3.17 *Resistances connected in parallel*

the battery. Since the potential difference across each resistance is the same and since each resistance has a different value, Ohm's law tells us that the current in each resistance should be different and must be as follows:

$$I_1 = \frac{V}{R_1} \qquad I_2 = \frac{V}{R_2} \qquad I_3 = \frac{V}{R_3}$$

Thus

$$I = V\left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}\right) \tag{3.24}$$

Now suppose that the resistance of the combination is R. This implies that if the parallel combination of resistances R_1, R_2 and R_3 is replaced by its equivalent resistance R, the total current through R and the potential difference across R must be the same.

$$\therefore \qquad I = \frac{V}{R}$$

Using this value of I in Eq. (3.24) we have

or

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}$$

For n resistances connected in parallel, we have

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \ldots + \frac{1}{R_n} \tag{3.25}$$

Thus *when a number of resistances are connected in parallel, the reciprocal of the resistance of the combination is equal to the sum of the reciprocals of the individual resistances.*

EXAMPLE 3.8 Two identical cells of emf 1.5 V each connected in parallel provide supply to an external circuit consisting of two resistors of 17 Ω each joined in parallel. A very high resistance voltmeter reads the terminal voltage of the cells to be 1.4 V. Find the internal resistance of each cell.

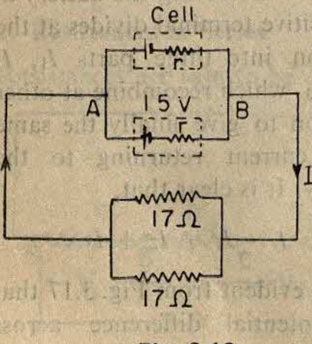


Fig. 3.18

Solution: Figure 3.18 shows the circuit diagram. I is the current in the circuit. The equivalent resistance R of the parallel combination of two 17 Ω resistors is given by

$$\frac{1}{R} = \frac{1}{17} + \frac{1}{17}$$

or
$$R = \frac{17}{2}\,\Omega$$

Now the terminal voltage V across points A and B = 1.4 V. Hence the current in the circuit is

$$I = \frac{V}{R} = \frac{1.4}{17/2} = \frac{2.8}{17}\,\text{A}$$

Let r be the internal resistance of each cell. The voltage drop across r is $= 1.5 - 1.4 = 0.1$ V. Hence, we have

$$\frac{2.8r}{17} = 0.1$$

giving
$$r = 1.2\,\Omega$$

EXAMPLE 3.9 A resistance network shown in Fig. 3.19 is connected to a battery of emf 30 V having an internal resistance of 1 Ω. Find: (a) the equivalent resistance of the network, (b) the current in each resistor and the potential drop across each resistor, and (c) the terminal voltage of the battery.

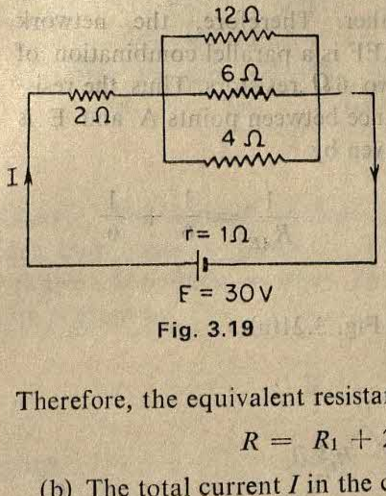


Fig. 3.19

Solution: (a) The total resistance of the parallel combination is given by

$$\frac{1}{R_1} = \frac{1}{12} + \frac{1}{6} + \frac{1}{4}$$

giving $R_1 = 2\,\Omega$

Therefore, the equivalent resistance of the network is

$$R = R_1 + 2 = 2 + 2 = 4\,\Omega$$

(b) The total current I in the circuit is

$$I = \frac{\mathcal{E}}{R + r} = \frac{30}{4 + 1} = 6\text{ A}$$

∴ Potential drop across the $2\,\Omega$ resistor $= 2 \times 6 = 12$ V

Now the parallel combination of $12\,\Omega$, $6\,\Omega$ and $4\,\Omega$ resistors has an effective resistance of $R_1 = 2\,\Omega$.

Hence the potential drop across the parallel combination $= 2 \times 6 = 12$ V which is the same for each resistor in the combination. Therefore

$$\text{Current in } 12\,\Omega \text{ resistor} = \frac{12\text{ V}}{12\,\Omega} = 1\text{ A}$$

$$\text{Current in } 6\,\Omega \text{ resistor} = \frac{12\text{ V}}{6\,\Omega} = 2\text{ A}$$

$$\text{Current in } 4\,\Omega \text{ resistor} = \frac{12\text{ V}}{4\,\Omega} = 3\text{ A}$$

(c) Total voltage drop across the network $= 12\text{ V} + 12\text{ V}$

$$= 24\text{ V}$$

Thus the terminal voltage of the battery is 24 V. The 'loss' of voltage $= 30\text{ V} - 24\text{ V} = 6\text{ V}$ is the potential drop across the internal resistance of the battery.

EXAMPLE 3.10 Find the effective resistance between points A and B of the network shown in Fig. 3.20.

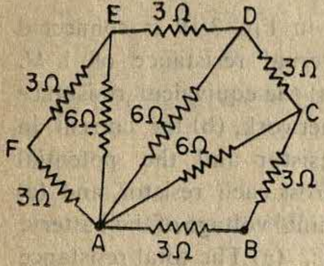


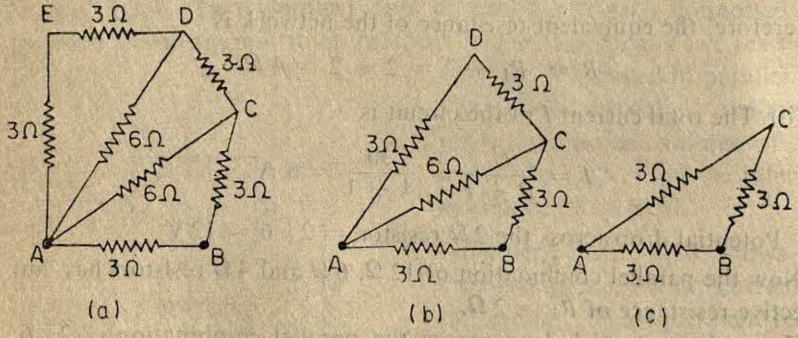
Fig. 3.20

Solution: Resistors AF and FE of 3 Ω each are in series with each other. Therefore, the network AEF is a parallel combination of two 6 Ω resistors. Thus the resistance between points A and E is given by

$$\frac{1}{R_{AE}} = \frac{1}{6} + \frac{1}{6}$$

giving $R_{AE} = 3\,\Omega$

The network reduces to that shown in Fig. 3.21(a).



(a) (b) (c)

Fig. 3.21

Similarly the resistance between points A and D in Fig. 3.21(a) is

$$R_{AD} = 3\,\Omega$$

The network reduces to that shown in Fig. 3.21(b). It follows from this figure that the resistance between points A and C is

$$R_{AC} = 3\ \Omega$$

The network simplifies to that shown in Fig. 3.21(c). Clearly the effective resistance between points A and B is given by

$$\frac{1}{R_{AB}} = \frac{1}{6} + \frac{1}{3}$$

or $$R_{AB} = 2\ \Omega$$

EXAMPLE 3.11 Find the equivalent resistance of the network shown in Fig. 3.22. The numbers indicate resistances in ohms.

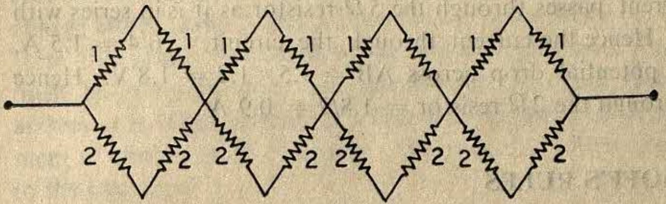


Fig. 3.22

Solution: The network consists of four loops connected in series. Each loop consists of two 1 Ω resistors in series connected in parallel with two 2 Ω resistors in series. Thus the effective resistance R of each loop is given by

$$\frac{1}{R} = \frac{1}{2} + \frac{1}{4}$$

giving $$R = \frac{4}{3}\ \Omega$$

∴ Equivalent resistance of the network

$= $ resistance of four loops connected in series, each having an effective resistance of $\frac{4}{3}\ \Omega$

$$= 4 \times \frac{4}{3} = \frac{16}{3}\ \Omega$$

EXAMPLE 3.12 Calculate the steady state current in the 2 Ω resistor shown in Fig. 3.23. The internal resistance of the battery is negligible and the capacitance of the capacitor is 0.5 μF.

Solution: The resistance of the parallel combination of 2 Ω and 3 Ω resistors is

$$R = 1.2\,\Omega$$

This resistance is in series with the 2.8 Ω resistor, giving a total effective resistance

$$= 1.2 + 2.8 = 4.0\,\Omega.$$

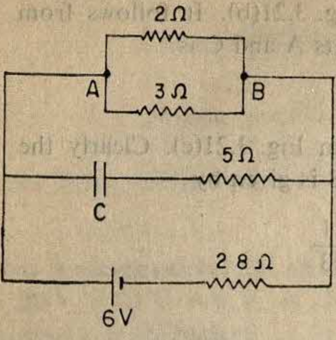


Fig. 3.23

In the steady state, no current flows through the capacitor C and hence no current passes through the 5 Ω resistor as it is in series with the capacitor. Hence the current through the circuit = 6/4 = 1.5 A. Therefore the potential drop across AB = 1.5×1.2 = 1.8 V. Hence the current through the 2 Ω resistor = 1.8/2 = 0.9 A.

3.8 KIRCHHOFF'S RULES

When we have to deal with complicated circuits (such as those we have in a radio receiver or a television set) which cannot be reduced to simple series and parallel arrangements, we can determine the equivalent resistance of the circuit or the current in any part of the circuit or the potential difference across any resistor in the circuit by using *Kirchhoff's rules* described below. These rules are generalizations based on Ohm's law.

Let us consider the circuit shown in Fig. 3.24. It consists of two *loops* abefa and bcdeb labelled 1 and 2. There are two *junctions* b and e and three branches connecting these junctions. The branches are the left branch bafe, the right branch bcde and the central branch be.

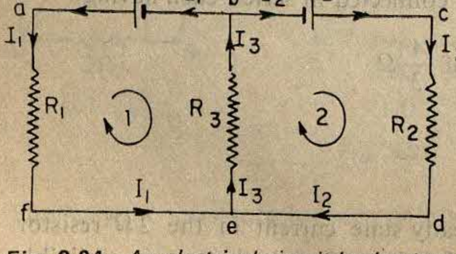


Fig. 3.24 *An electrical circuit having two loops*

There are two batteries of emfs $\mathcal{E}_1$ and $\mathcal{E}_2$. For simplicity, we neglect their internal resistances. The currents in the three branches are I_1, I_2 and I_3 as shown. Current I_1 has the same value at any point on the left branch, I_2 has the same value everywhere in the right branch and I_3 in the central branch.

If the emfs and the resistances are known, the currents in the various branches can be found very simply by using the following two rules.

1. First or Junction rule

This rule is based on the fact that charge does not accumulate at any point in a conductor. It states that at any junction of several circuit elements, *the sum of the currents entering the junction must equal the sum of the currents leaving it.* The three currents I_1, I_2 and I_3 carry charge either towards the junction b or away from it. In the steady state (i.e. when the currents are not changing with time), the charge must be removed from the junction at the same rate at which it arrives at it. Thus Kirchhoff's first (or junction) rule is simply a statement of conservation of charge. Applying this rule to junction b leads to the equation

$$I_1 + I_2 = I_3 \tag{3.26}$$

Applying this rule to the other junction e leads to exactly the same equation. To solve for the three unknowns I_1, I_2 and I_3 we need two more independent equations. They can be found from Kirchhoff's second (or loop) rule.

2. Second or Loop Rule

In any closed loop, the algebraic sum of the changes in potential across the sources of current and resistances must be zero. The sources of current are the emfs and the changes in potential across the resistances are the voltage drops.

In order to deal with complicated circuits, we must fix the sign conventions for both the emf and the voltage drops in resistance. In using the second rule, it is necessary to start at some point in the loop and move completely around the loop in one direction or the other, keeping track of sign of each emf ($\mathcal{E}$) and each voltage drop (IR) through each resistor. The sign of each emf and each IR voltage drop is related to the direction we choose to go around the loop. In the circuit shown in Fig. 3.24 suppose we decide to go around a loop in the clock-wise direction abefa for loop 1 and bcdeb for loop 2. *If the direction of the path taken to go round the loop is the same as the* *direction in which the positive charge (or positive current) moves in the source (battery), the emf of the source is taken to be positive. The sign of the IR drop in each resistor in the loop is positive when the direction of the travel is the same as the direction of the current.*

With these sign conventions, using Kirchhoff's second rule to loop abefa we have

$$+\mathcal{E}_1 - I_3R_3 - I_1R_1 = 0 \tag{3.27}$$

For the second loop bcdeb we have

$$-\mathcal{E}_2 + I_2R_2 + I_3R_3 = 0 \tag{3.28}$$

The three equations [Eqs (3.26), (3.27) and (3.28)] are simultaneously solved for I_1, I_2 and I_3. Thus knowing $\mathcal{E}_1$, $\mathcal{E}_2$, R_1, R_2 and R_3, the currents I_1, I_2 and I_3 are determined and hence the potential drops across the resistors are known. There are three equations and exactly three unknowns.

Note: We could also apply the second rule to the large loop consisting of entire circuit abcdefa. For this loop, we have

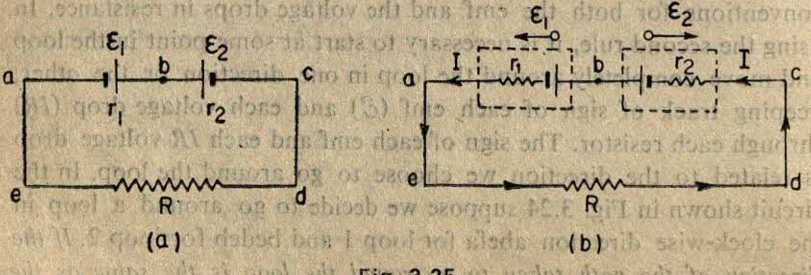
$$\mathcal{E}_1 - \mathcal{E}_2 + I_2R_2 - I_1R_1 = 0 \tag{3.29}$$

Notice that this is not a new independent equation, it is simply the sum of Eqs (3.27) and (3.28).

The following examples illustrate the use of Kirchhoff's rules to solve electrical circuit problems.

EXAMPLE 3.13 Two cells of emfs $\mathcal{E}_1 = 1.5$ V and $\mathcal{E}_2 = 2.0$ V and internal resistances of $r_1 = 1.0\ \Omega$ and $r_2 = 1.5\ \Omega$ respectively are connected to an external resistor $R = 2.5\ \Omega$ as shown in Fig. 3.25(a). Find: (a) the current in the circuit and (b) the potential difference between points a and b and between b and c.

Solution: (a) Let us suppose that the direction of the current I is as shown in Fig. 3.25 (b); this direction is chosen arbitrarily. It is not necessary to know in advance the direction of the current in any



(a) (b)

Fig. 3.25

branch in any complicated circuit. We take the current in any direction and apply Kirchhoff's rules to the circuit. If the numerical value of the current turns out to be positive, then the chosen direction is correct; but if the value turns out to be negative, the actual direction of the current is opposite to the one chosen.

Applying the loop rule to the loop abcdea, we have

$$-Ir_1 - \mathcal{E}_1 + \mathcal{E}_2 - Ir_2 - IR = 0$$

or

$$I = \frac{\mathcal{E}_2 - \mathcal{E}_1}{R + r_1 + r_2}$$

$$= \frac{2.0 - 1.5}{2.5 + 1.0 + 1.5} = 0.10 \text{ A}$$

which is positive. Hence the direction chosen for I is correct

(b) To find the potential difference between points a and b we start from a and traverse the circuit to b, obtaining

$$V_{ab}(=V_a - V_b) = -Ir_1 - \mathcal{E}_1 = -0.10 \times 1.0 - 1.5 = -1.6 \text{ V}$$

The negative sign indicates that point a is at a lower potential than point b.

The potential difference between points b and c is given by

$$V_{bc}(=V_b - V_c) = \mathcal{E}_2 - Ir_2 = 2.0 - 0.10 \times 1.5 = +1.85 \text{ V}$$

The positive sign shows that point b is at a higher potential than point c. Notice that the terminal voltage of $\mathcal{E}_1$ is 1.6 V which is *greater than* its emf (which is 1.5 V) and the terminal voltage of $\mathcal{E}_2$ is 1.85 V which is *less than* its emf(which is 2.0 V). This means that charge is being forced through $\mathcal{E}_1$ in a direction opposite to the one in which it would send charge if it were acting by itself. If $\mathcal{E}_1$ is a chargeable battery (storage battery), it is being charged at the expense of $\mathcal{E}_2$.

EXAMPLE 3.14 In the circuit shown in Fig. 3.26, $\mathcal{E}_1 = 3$ V, $\mathcal{E}_2 = 2$ V, $\mathcal{E}_3 = 1$ V and $R = r_1 = r_2 = r_3 = 1\ \Omega$. (a) Find the current through

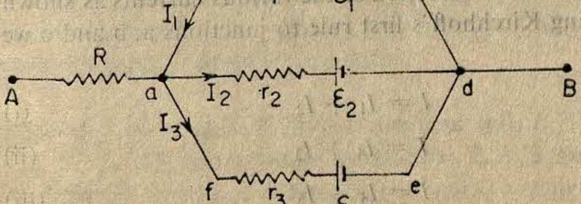


Fig. 3.26

each branch. (b) If r_2 is short-circuited and the point A is connected to point B, find the the currents through the cells $\mathcal{E}_1$, $\mathcal{E}_2$, $\mathcal{E}_3$ and the resistor R.

Solution: (a) Let I_1, I_2 and I_3 be the currents through the resistors r_1, r_2, and r_3 respectively as indicated in Fig. 3.26. Using Kirchhoff's second rule to loops abcda and abcdefa, we have

$$-I_1r_1 + \mathcal{E}_1 - \mathcal{E}_2 - I_2r_2 = 0$$

and $$-I_1r_1 + \mathcal{E}_1 - \mathcal{E}_3 - I_3r_3 = 0$$

which give $$\mathcal{E}_1 - I_1r_1 = \mathcal{E}_2 + I_2r_2 = \mathcal{E}_3 + I_3r_3 \qquad \text{(i)}$$

Applying Kirchhoff's first rule to junction a, we have

$$I_1 = I_2 + I_3 \qquad \text{(ii)}$$

Using (ii) in (i), we get

$$\mathcal{E}_1 - (I_2 + I_3)r_1 = \mathcal{E}_3 + I_3r_3$$

or $$2I_3 + I_2 = 2 \qquad \text{(iii)}$$

Also $$\mathcal{E}_2 + I_2r_2 = \mathcal{E}_3 + I_3r_3$$

or $$I_3 - I_2 = 1 \qquad \text{(iv)}$$

From (iii) and (iv), we get

$$I_1 = 1\ \text{A}, \qquad I_2 = 0 \quad \text{and} \quad I_3 = 1\ \text{A}$$

(b) Since $I_2 = 0$, the potential difference between points a and d = emf $\mathcal{E}_2 = 2$ V and remains equal to 2 V even when r_2 is short-circuited. Because the potential difference across a and d remains unchanged, the currents I_1 and I_3 through cells $\mathcal{E}_1$ and $\mathcal{E}_2$ do not change i.e. $I_1 = I_3 = 1$ A, even when point A is connected to point B. Hence the current through $\mathcal{E}_2$ will be 2 A.

EXAMPLE 3.15 Determine the current in each branch of the net work shown in Fig. 3.27.

Solution: Let us choose the direction of the various currents as shown in Fig. 3.27. Applying Kirchhoff's first rule to junctions a, b and c we have

For junction a: $$I = I_1 + I_2 \qquad \text{(i)}$$

For junction b: $$I = I_4 + I_3 \qquad \text{(ii)}$$

For junction c: $$I = I_4 + I_5 \qquad \text{(iii)}$$

For junction d: $$I_5 = I_3 + I_2 \qquad \text{(iv)}$$

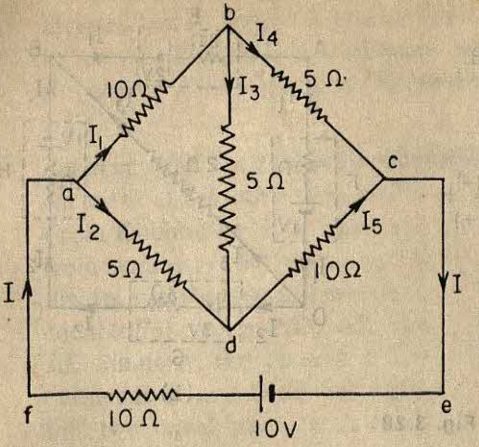


Fig. 3.27

Applying Kirchhoff's second rule to loops abde, bcdb and adcefa we have

$$10I_1 + 5I_3 - 5I_2 = 0 \quad \text{or} \quad 2I_1 + I_3 - I_2 = 0 \tag{v}$$

$$5I_4 - 10I_5 - 5I_3 = 0 \quad \text{or} \quad I_4 - 2I_5 - I_3 = 0 \tag{vi}$$

$$5I_2 + 10I_5 - 10 + 10I = 0 \quad \text{or} \quad I_2 + 2I_5 - 2 + 2I = 0 \tag{vii}$$

Eliminating I between (i) and (vii) we get

$$3I_2 + 2I_5 - 2 + 2I_1 = 0 \tag{viii}$$

We have five equations (ii), (iv), (v), (vi) and (viii) and five unknowns I_1, I_2, I_3, I_4 and I_5. Solving these five equations simultaneously we get

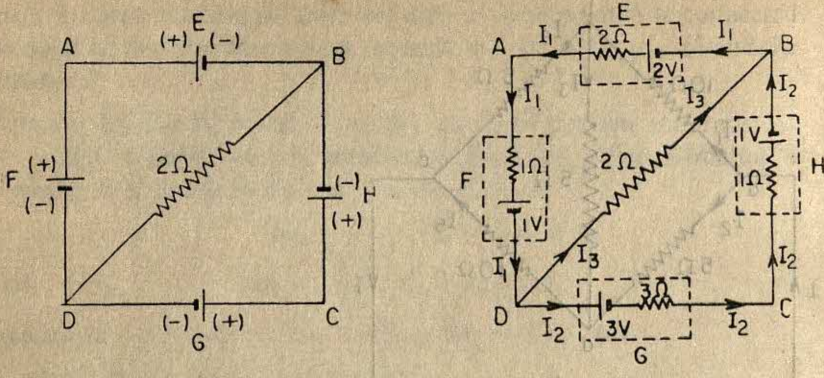
$$I_1 = \frac{4}{17}\text{ A}, \quad I_2 = \frac{6}{17}\text{ A}, \quad I_3 = -\frac{2}{17}\text{ A}, \quad I_4 = \frac{6}{17}\text{ A}$$

and
$$I_5 = \frac{4}{17}\text{ A}.$$

Hence
$$I = I_1 + I_2 = \frac{10}{17}\text{ A}$$

Since I_3 turns out to be negative, the direction I_3 chosen as shown in the figure is wrong, I_3 flows from d to b.

EXAMPLE 3.16 Figure 3.28(a) shows four cells E, F, G and H of emfs 2, 1, 3 and IV and internal resistances 2, 1, 3 and 1 Ω respectively. Calculate the potential difference between: (a) points B and D and (b) the terminals of cells G and H.



(a) (b)

Fig. 3.28

Solution: Let I_1 and I_2 be the currents in branches BAD and DCB respectively as shown in Fig. 3.28(b). Let I_3 be the current along DB through the 2 Ω resistor. Applying Kirchhoff's first law at junction D, we have

$$I_1 = I_2 + I_3$$

or
$$I_3 = I_1 - I_2$$

Applying Kirchhoff's second law to loops BADB and DCBD we have

$$2I_1 + I_1 + 2(I_1 - I_2) = 2 - 1 = 1$$

or
$$5I_1 - 2I_2 = 1 \qquad \text{(i)}$$

and
$$3I_2 + I_2 - 2(I_1 - I_2) = 3 - 1 = 2$$

or
$$6I_2 - 2I_1 = 2 \qquad \text{(ii)}$$

From (i) and (ii), we get

$$I_1 = \frac{6}{13}\text{ A} \qquad \text{and} \qquad I_2 = \frac{5}{13}\text{ A}.$$

Therefore
$$I_3 = I_1 - I_2 = -\frac{1}{13}\text{ A}$$

The negative sign shows that the current through the 2 Ω resistor flows along BD and not along DB as assumed.

(a) Potential difference between B and D $= 2\times(1/13) = 2/13$ V

(b) Potential difference across the terminals of cell G

$$= 3 - (6/13)\times 3 = 21/13\text{V}$$

Potential difference across the terminals of cell H

$$= 1 + 6/13 = 9/13\text{ V}.$$

EXAMPLE 3.17 Twelve identical wires, each of resistance r, are joined to form a cube. Find the resistance between (a) two diagonally opposite corners of the cube and (b) two corners on the same edge of the cube.

Solution: (a) Let I be the current entering at corner A and leaving at corner G of the cube (Fig. 3.29). From Kirchhoff's first rule and from symmetry, the current at A divides into three equal parts (each equal to $I/3$) along AD, AB and AE. Similarly, the current $I/3$ at corners B, E and D divides again into two equal parts (each equal to $I/6$). Thus the currents in all the twelve resistances are known; they are shown in Fig. 3.29.

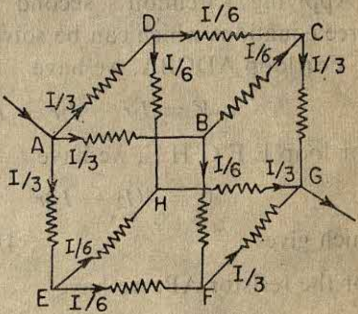


Fig. 3.29

Let V be the potential difference between points A and G. Applying Kirchhoff's second rule to any one path (such as A E F G), we have

$$V = \frac{Ir}{3} + \frac{Ir}{6} + \frac{Ir}{3} = \frac{5Ir}{6}$$

From symmetry, it follows that the same result is obtained for all paths between A and G. If R_{AG} is the effective resistance between points A and G, we have

$$V = IR_{AG} = \frac{5Ir}{6}$$

giving

$$R_{AG} = \frac{5r}{6}$$

(b) Let us consider two corners A and B of the same edge of the cube, as shown in Fig. 3.30. Let V be the potential difference between A and B. Let current I enter at A and leave at B.

Suppose the current in the resistor AB is I_1. Let I_2 be the current in resistor AE, which is the same as that in resistor AD. Also let I_3 be the current in resistor DC which is the same as that in resistor EF.

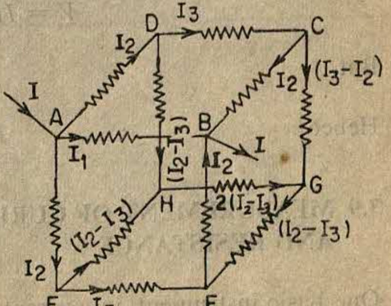


Fig. 3.30

From Kirchhoff's first rule, the currents in DH and EH are each equal to $(I_2 - I_3)$. Similarly, the current in HG is $2(I_2 - I_3)$. The currents in CB and BF are each equal to I_2. The current in CG is $(I_3 - I_2)$ and that in GF is $(I_3 - I_2) + 2(I_2 - I_3) = (I_2 - I_3)$. Thus the currents in all the twelve resistors are specified.

Applying Kirchhoff's second rule to any three loops, we obtain three equations which can be solved for the three unknowns I_1, I_2 and I_3. For loop ADCBA, we have

$$V = I_2r + I_2r + I_3r = (2I_2 + I_3)r \tag{i}$$

For loop E F G H E, we have

$$0 = 4(I_2 - I_3)r - I_3r = (4I_2 - 5I_3)r$$

which gives

$$4I_2 = 5I_3 \tag{ii}$$

For the resistor AB, we have

$$V = I_1r \tag{iii}$$

Using (iii) in (i), we get

$$V = \left(2I_2 + \frac{4}{5}I_2\right)r = \frac{14}{5}I_2r \tag{iv}$$

Using (iii) in (iv), we get

$$I_1 = \frac{14}{5}I_2$$

The total current is

$$I = I_1 + I_2 + I_2 = I_1 + \frac{10}{14}I_1 = \frac{12}{7}I_1$$

If R_{AB} is the effective resistance of the whole network between points A and B, we have

$$V = IR_{AB} = \frac{12}{7}I_1R_{AB}$$

But

$$V = I_1r$$

Hence

$$R_{AB} = \frac{7}{12}r$$

3.9 MEASUREMENT OF CURRENT, POTENTIAL DIFFERENCE AND RESISTANCE

One basic instrument in current electricity is the galvanometer. It consists of a coil of many turns of an insulated wire suspended or pivoted between the two pieces of a strong permanent horse-shoe magnet. The working of a galvanometer will be discussed in Chapter 5.

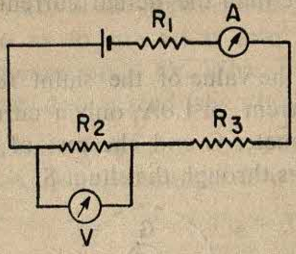


Fig. 3.31 *An ammeter connected to read the current in the circuit and a voltmeter connected to read the potential difference across resistor R_2*

When a current is passed through the coil, it experiences a torque and deflects. The deflection of the coil is proportional to the current flowing through the coil. A galvanometer can be easily converted into devices for measuring currents and potential differences called ammeters and voltmeters. The requirement is that the use of these devices in a circuit must *not* alter the current in the circuit or the potential difference across any resistor in the circuit. One such circuit is shown in Fig. 3.31 where A stands for an ammeter and V for a voltmeter.

The Ammeter

An ordinary moving coil galvanometer is a very sensitive instrument. It is designed to read currents as small as a few microamperes. To measure currents as high as 1 A or more, a low resistance is connected in parallel with the galvanometer coil. The resistance used in this way is called a *shunt. A shunted galvanometer is called an ammeter.* Since the resistance of the shunt is very small compared to that of the coil, most of the current will flow through the shunt and only a small fraction passes through the coil.

An ammeter is always connected in *series* with the circuit the current through which is to be measured as shown in Fig. 3.31. It is essential that the resistance r of the ammeter be *small* compared to other resistances in the circuit. In the circuit shown in Fig. 3.31, the requirement is that

$$r \ll (R_1 + R_2 + R_3)$$

Otherwise the very act of connecting the ammeter will itself alter the current to be measured. An ideal ammeter would have a *zero* resistance. The value of the shunt resistance determines the range of the currents a given ammeter can measure (see Example 3.18).

EXAMPLE 3.18 A galvanometer of resistance 20 Ω gives full-scale deflection when a current of 0.001 A is passed through it. Calculate the value of the shunt that must be used to convert it into an ammeter than can read currents up to 1.0 A. When this ammeter is put in a circuit, does it read (slightly) less or more than the actual current in the original circuit?

Solution: It is clear from Fig. 3.32 that the value of the shunt resistance must be such that out of a current of 1.0A, only a current $I_G = 0.001$A passes through the galvanometer G and the rest of the current $I_S = 1.0 - 0.001 = 0.999$A passes through the shunt S.

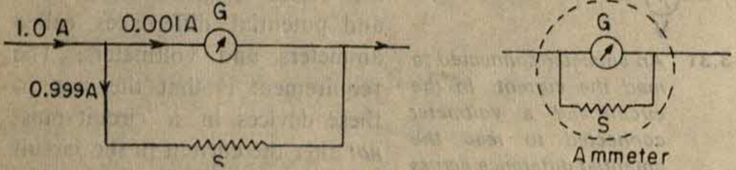


Fig. 3.32 *Converting a galvanometer into an ammeter*

Let the resistance of the shunt be R_S and that of the galvanometer R_G. Since R_G and R_S are in parallel, the potential difference across each is the same, say V. From Ohm's law, we have

$$V = I_G \times R_G = I_S \times R_S$$

or

$$0.001 \times 20 = 0.999 \times R_S$$

or

$$R_S = 0.020\,\Omega$$

Thus, the given galvanometer shunted with a resistance of 0.020 Ω becomes an ammeter which will give a full-scale deflection for a current of 1.0A. Since $R_S \ll R_G$, the effective resistance of the ammeter $\simeq R_S = 0.020\ \Omega$. When the ammeter is put in a circuit, it will decrease (slightly) the original current. Hence it will read slightly less than the actual current. See example 3.20.

The Voltmeter

A voltmeter is designed to measure potential differences across the resistors in a circuit. *A voltmeter is a galvanometer with a high resistance connected in series with the coil.* It is always connected in *parallel* to the resistor the potential difference across which is to be measured as shown in Fig. 3.31. The voltmeter must have a *high* resistance (R) compared to the resistance (R_2) across which the potential difference is measured. The voltmeter will then take a negligible current so that the current (and hence the potential difference across any resistor) in the main circuit remains practically unaffected when the voltmeter is connected. The range of the voltmeter is determined by the value of the high resistance R (see Example 3.19).

EXAMPLE 3.19 A galvanometer of resistance 20 Ω gives full-scale deflection when a current of 0.001 A is passed through it. Calculate the value of the high resistance that must be connected in series with it so as to convert it into a voltmeter that can measure potential differences up to 5V. When this voltmeter is put across a part of the circuit, does it read (slightly) less or more than the original voltage?

Solution: It is clear from Fig. 3.33 that the potential difference V_G across the galvanometer G when it gives full-scale deflection is given by

$$V_G = \text{Current} \times \text{Resistance}$$
$$= 0.001 \times 20 = 0.02 \text{ V}$$

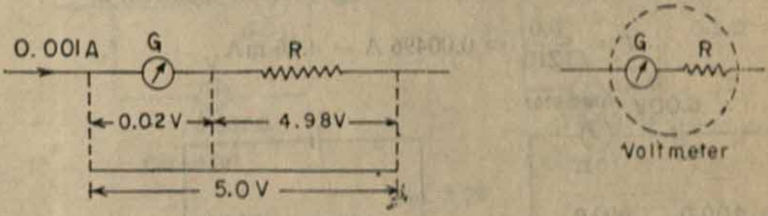


Fig. 3.33 *Conversion of a galvanometer into a voltmeter*

The value of resistance R that must be connected in series with the galvanometer of resistance R_G must be such that out of a total potential difference of 5 V only 0.02 V $= V_G$ exists across G and the rest $V_R = 5 - 0.02 = 4.98$ V drops across R. Since the current is the same in both G and R, namely, equal to 0.001 A. hence

$$R = \frac{V_R}{I} = \frac{4.98}{0.001} = 4980\,\Omega$$

Thus, the galvanometer with a resistance of 4980 Ω in series with it becomes a voltmeter which will give full-scale deflection when a potential difference of 5V is applied across the combination.

When this voltmeter whose effective resistance is (~4980 Ω) is connected across a resistor $r(\ll R)$, the current in the main circuit divides between r and R, the voltmeter taking a very small current. Hence the current through r decreases slightly. Consequently, the voltmeter reads slightly less than the actual voltage across r. See Example 3.20.

EXAMPLE 3.20 Two resistors of 400 Ω and 800 Ω are connected in series with a 6.00 V battery of negligible internal resistance. (a) An ammeter of 10.0 Ω resistance is used to measure the current in the circuit. Find the error in the measurement of current. (b) A voltmeter of resistance

10,000 Ω is used to measure the potential difference across the 400 Ω resistor. Find the error in the measurement of potential difference.

Solution: The current in the circuit is

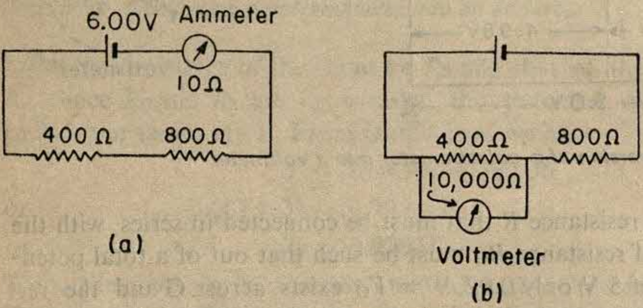
$$I = \frac{6.00}{400 + 800} = 0.005 \text{ A} = 5.00 \text{ mA}$$

The potential difference across the 400 Ω resistor is

$$V = 0.005 \times 400 = 2.00 \text{ V}$$

(a) When an ammeter of resistance 10 Ω is connected in series as shown in Fig. 3.34(a), the total resistance of the circuit is 400 + 800 + 10 = 1210 Ω. Therefore the ammeter will indicate a current of

$$I' = \frac{6.0}{1210} = 0.00496 \text{ A} = 4.96 \text{ mA}$$



(a)

(b)

Fig. 3.34

Hence the ammeter reads less than the actual current which is 5.00 mA and the error in the measurement of current is 5.00 − 4.96 = 0.04 mA.

(b) When a voltmeter of resistance 10,000 Ω is connected across the 400 Ω resistor as shown in Fig. 3.34(b), the equivalent resistance R of the combination is given by

$$\frac{1}{R} = \frac{1}{10{,}000} + \frac{1}{400}$$

which gives $R = 384.6\,\Omega$

Total resistance in the circuit = 800 + 384.6 = 1184.6 Ω

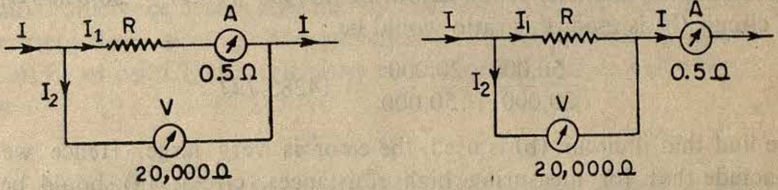
$$\text{Current in the circuit} = \frac{6.00}{1184.6} = 5.065 \times 10^{-3} \text{ A}$$

The potential difference indicated by the voltmeter is

$$V' = 384.6 \times 5.065 \times 10^{-3} = 1.95 \text{ V}$$

Hence the voltmeter reads less than the actual potential difference V(which is 2.00 V) and the error in the measurement of potential difference is $2.00 - 1.95 = 0.05$ V.

EXAMPLE 3.21 You are given two resistors X and Y whose resistances are to be determined using an ammeter of resistance 0.5 ohms and a voltmeter of resistance 20 kΩ. It is known that X is in the range of a few ohms and Y is in the range of several thousand ohms. In each case, which of the two circuits shown in Fig. 3.35 would you choose for resistance measurement? Justify your answer quantitatively.



Circuit (a)

Circuit (b)

Fig. 3.35

Solution: Let us find the error in the measurement of X and Y using circuits (a) and (b) of Fig. 3.35. The circuit that corresponds to a small error should be chosen.

(1) *Measurement of X*: Let us suppose that the resistance X is 5.0 Ω and find the ratio of the voltmeter and ammeter readings using circuits (a) and (b). The circuit for which this ratio is closer to 5.0 Ω is to be preferred.

Referring to circuit (a) the current I divides into I_1 and I_2 so that $I = I_1 + I_2$. It is clear that the ammeter reading is I_1, and the voltmeter reading $= I_1 \times (R + 0.5)$

$$= I_1 \times (5.5)$$

since, in this case $R = X = 5.0\,\Omega$. Hence the ratio

$$\frac{\text{Voltmeter reading}}{\text{Ammeter reading}} = 5.5\ \Omega$$

Thus, if we use circuit (a), the error in the measurement of X is 0.5 Ω in 5.0 Ω.

Referring to circuit (b), the ammeter reading is $I(=I_1 + I_2)$ and th voltmeter reading $= I_1 R = I_1 X = 5.0 I_1$. But $5.0 I_1 = 20{,}000\ I_2$. Thu we have

$$I_1 = \frac{20{,}000}{20{,}005}\, I$$

$$\text{Voltmeter reading} = 5.0\,I_1 = \frac{5.0 \times 20{,}000}{20{,}005}\,I$$

Hence $$\frac{\text{Voltmeter reading}}{\text{Ammeter reading}} = \frac{5.0 \times 20{,}000}{20{,}005} = 4.9987\ \Omega$$

We find that if we use circuit (b) for the measurement of a resistance in the range of a few ohms, the error in the measurement is negligibly small compared to that in circuit (a). Hence circuit (b) should be used.

(2) *Measurement of Y*: Resistance Y is several thousand ohms. Let us assume it to be, say, 50,000 Ω. If we use circuit (a), the ratio of the voltmeter and ammeter readings will be $(50{,}000 + 0.5) = 50{,}000.5\ \Omega$. If circuit (b) is used, this ratio would be

$$\frac{50{,}000 \times 20{,}000}{20{,}000 + 50{,}000} = 14285.7\ \Omega$$

We find that if circuit (b) is used, the error is very large. Hence we conclude that for measuring high resistances, circuit (a) should be used.

The Potentiometer

The potentiometer is a device used for accurate comparison of potentials differences. It is schematically shown in Fig. 3.36.

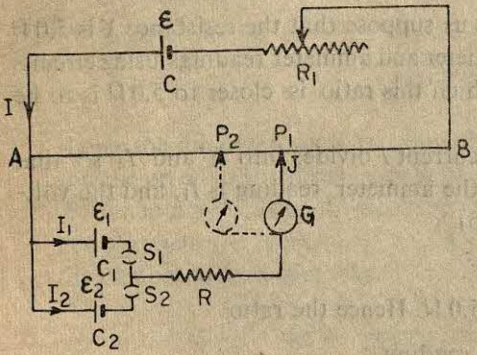


Fig. 3.36 *The circuit diagram of a potentiometer. At null point with* C_1, $I_1 = 0$ *and with* C_2, $I_2 = 0$

A source (cell C) of emf $\mathcal{E}$ is connected to a variable resistor R_1 (called the rheostat) and a long uniform conducting wire AB in series as shown in Fig. 3.36. Two cells, C_1 and C_2 whose emfs $\mathcal{E}_1$ and $\mathcal{E}_2$ are to be compared, are connected as shown (with the positive terminals of all the three cells connected to A). S_1 and S_2 are two switches.

They are used to bring either C_1 or C_2 in the circuit. G is a zero-centre galvanometer that detects the current flowing through it. A high resistance R is connected in series with G to prevent too much current from flowing through G. The other end of G is connected to a jockey J which slides over the wire AB.

First the switch S_1 is closed so that the cell C_1 is brought into the circuit. The jockey J is moved along the wire AB till the current in the galvanometer is zero and it shows no deflection. Suppose the null point occurs when J is at point P_1. This happens if the potential difference (V_1) across the length AP_1 of the wire due to cell C = the potential difference across AP_1 due to cell C_1. Since, at null point, no current flows through the galvanometer circuit, the latter equals the emf $\mathcal{E}_1$ of cell C_1, this cell being on an open circuit at null point. Thus

$$\mathcal{E}_1 = V_1$$

Let V be the terminal voltage of cell C and k, the resistance per unit length of wire AB. Then, from Ohm's law, it follows that the potential difference V_1 across AP_1 is given by

$$V_1 = \frac{V\,k\mathrm{AP}_1}{(R_1 + k\mathrm{AB})} = Ik\,\mathrm{AP}_1$$

where $I = V/(R_1 + k\mathrm{AB})$ is the current in the circuit involving AB, C and R_1. Therefore

$$\mathcal{E}_1 = Ik\mathrm{AP}_1 \tag{3.30}$$

Next the switch S_2 is closed (switch S_1 being open) so that the cell C_2 is brought into the circuit. Suppose now the null point is at P_2. Then, the emf $\mathcal{E}_2$ of the cell C_2 is given by

$$\mathcal{E}_2 = Ik\mathrm{AP}_2 \tag{3.31}$$

Dividing Eqs. (3.30) and (3.31) we get

$$\frac{\mathcal{E}_1}{\mathcal{E}_2} = \frac{\mathrm{AP}_1}{\mathrm{AP}_2} \tag{3.32}$$

Notice that, at null point, the current (I) flows only in the circuit involving wire AB, resistor R_1 and cell C and no current flows through either C_1 or C_2 ($I_1 = I_2 = 0$). Therefore, the internal resistances of cells C_1 and C_2 do not play any role. Hence the emfs (and not the terminal voltages) of C_1 and C_2 are compared by measuring the lengths AP_1 and AP_2 and finding their ratio [Eq. (3.32)]. If one of the emfs (say $\mathcal{E}_2$) is a standard known emf, this device can be used to find the absolute emf $\mathcal{E}_1$ of any source.

EXAMPLE 3.22 Figure 3.37 shows a potentiometer using a cell C of emf 2.0 V and internal resistance 0.40 Ω connected to a resistor wire AB. A standard cell of a constant emf of 1.02 V (for moderate currents up to a few amperes) gives a balance point at 67.3 cm length of the wire. A very high resistance $R = 600\ \mathrm{k}\Omega$ is put in series with the standard cell to ensure that low currents are drawn from the cell. This resistance is shorted by inserting switch S when close to the balance point. The standard cell is then replaced by a cell of unknown emf $\mathcal{E}$ and the balance point found by using the above procedure turns out to be 82.3 cm length of the wire. (a) What is the value of $\mathcal{E}$? (b) What is the purpose of using the high resistance R? (c) Is the balance point affected by this high resistance? (d) Is the balance point affected by the internal resistance of the driver cell C? (e) Would this method work if: (i) the internal resistance of cell C were higher than the resistance of wire AB and (ii) the emf of cell C were 1.0 V instead of 2.0 V? (f) Would the circuit work well for determining extremely small emfs of the order of a few millivolts? If not, how would you modify the circuit?

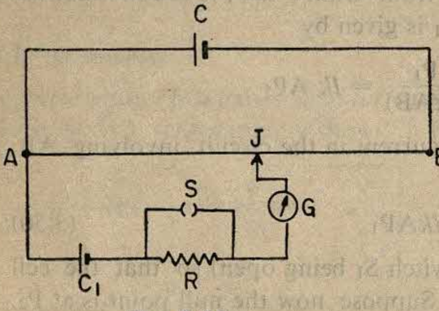


Fig. 3.37

Solution: (a) The value of $\mathcal{E}$ is given by ($\mathcal{E}_1 = 1.02$ V is the emf of the standard cell C_1)

$$\mathcal{E} = \frac{1.02 \times 82.3}{67.3} = 1.25\ \mathrm{V}$$

(b) The high resistance R has a two-fold purpose—to keep the current drawn from the standard cell within permissible limit when far away from the balance point and to prevent a large current to flow through the galvanometer (so that it is not damaged) when far away from the balance point.

(c) No, since at balance point no current flows through R.

(d) No, the balance point is not affected by the internal resistance of C, it depends on the terminal voltage of C and the emf of C_1.

(e) The method would not work if the potential difference across AB due to the driver cell C (which is its terminal voltage) became less than the emf of cell C_1, for then there would be no balance point on the wire. (i) If the internal resistance of C is higher than the resistance of AB, the potential difference across AB due to cell C of emf 2.0 V will be less than 1.0 V. Since the emf of the standard cell is greater than this value, there will be no balance point on the wire. (ii) For the same reason, the method would not work if the emf of C were 1.0 V instead of 2.0 V.

(f) Suppose the emf $\mathcal{E}$ to be measured is, say 5 mV $= 5\times10^{-3}$ V, and the potential difference across AB due to C is 2.0 V. Also suppose that the length AB of the wire is 100 cm and its resistance per centimetre is k. Then the value of the balance length will be given by

$$\begin{aligned} 5\times10^{-3}\ \text{V} &= \text{pd across AJ due to C} \\ &= \text{current in the main circuit} \times \text{resistance of AJ} \\ &= \frac{2.0}{100k}\cdot k\text{AJ} = \frac{2.0\ \text{AJ}}{100}\ \text{V} \end{aligned}$$

or

$$\text{AJ} = 0.25\ \text{cm}$$

Thus the balance point will be very close to A and there is an extremely large percentage error in its measurement. To have a large balance length, the potential difference across AB must be reduced to about 10 mV. To achieve this the circuit shown in Fig. 3.37 is modified by putting a suitable resistor R' as shown in Fig. 3.38. The balance length will then be larger and the percentage error will be much smaller.

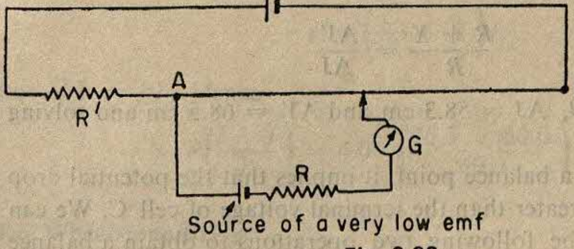


Fig. 3.38

EXAMPLE 3.23 Figure 3.39 shows a potentiometer circuit for comparing two resistances. The balance point with $R = 10.0\ \Omega$ when switch S_1 is closed and S_2 is open is found to be 58.3 cm while that when S_2 is closed (S_1 is open) is found to be 68.5 cm. Find the value of X. What will you do if you fail to find a balance point with the given cell C′.

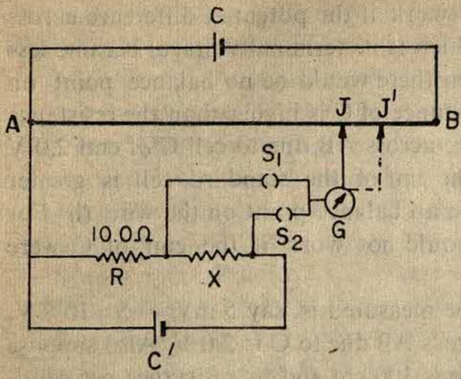


Fig. 3.39

Solution: Let V and V' be the terminal voltages of cells C and C' respectively. Let k be the resistance per unit length of wire AB, and J the balance point when S₁ is closed. This happens when the potential difference across AJ due to C = potential difference across R due to C', i.e.

$$\frac{V}{k\text{AB}} \cdot k\text{AJ} = \frac{V'R}{(R+X)} \tag{i}$$

Let J' be the balance point when switch S_1 is open and S_2 is closed, then

$$\frac{V}{k\text{AB}} \cdot k\text{AJ}' = V' \tag{ii}$$

Dividing (i) and (ii) we get

$$\frac{R+X}{R} = \frac{\text{AJ}'}{\text{AJ}}$$

Putting $R = 10.0\ \Omega$, AJ = 58.3 cm and AJ' = 68.5 cm and solving we get $r = 1.75\ \Omega$.

If we fail to find a balance point, it implies that the potential drop across $(R + X)$ is greater than the terminal voltage of cell C. We can perform either of the following two operations to obtain a balance point:

(i) Use one or more cells in series with C or (ii) reduce the current in $(R + X)$ and hence potential drop across $(R + X)$ by putting a suitable resistor in series with C'.

EXAMPLE 3.24 Figure 3.40 shows a potentiometer circuit for determining the internal resistance of a cell. When switch S is open, the balance point is found to be at 76.3 cm of the wire. When switch S is closed and the value of R is 4.0 Ω, the balance point shifts to 60.0 cm. Find the internal resistance of cell C′.

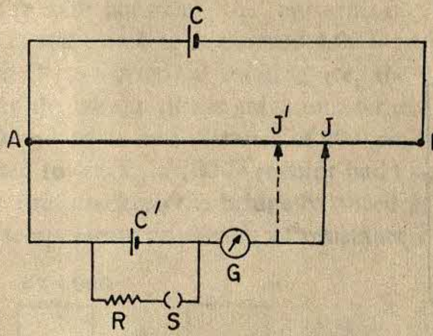


Fig. 3.40

Solution: Let $\mathcal{E}$ be the emf of the cell C′ and r its internal resistance. Let l = AJ be the balance length when switch S is open. When a resistance R is introduced by closing the switch, a current begins to flow through the cell C′ and resistance R. The potential difference between the terminals of the cell falls and the balance length decreases to l' = AJ′. The internal resistance of the cell is given by

$$r = \frac{\mathcal{E} - V}{I}$$

where V is the terminal voltage of C′ and I is the current in the circuit involving C′ and R. Also $I = V/R$. Hence

$$r = \left(\frac{\mathcal{E}}{V} - 1\right)R$$

But $\mathcal{E}/V = l/l'$. Hence

$$r = R\left(\frac{l - l'}{l'}\right) = 4.0 \times \left(\frac{76.3 - 60.0}{60.0}\right) = 1.1\ \Omega$$

The Wheatstone Bridge

The Wheatstone bridge is an arrangement of four resistances; three of them are known and the fourth can be determined in terms of the three known resistances. It was first suggested by C.F. Wheatstone of England in the 19th century. Figure 3.41 shows a Wheatstone bridge. It consists of three known resistances P, Q and S and a fourth unknown resistance R. A source of emf is connected between points A and C and a galvanometer between B and D.

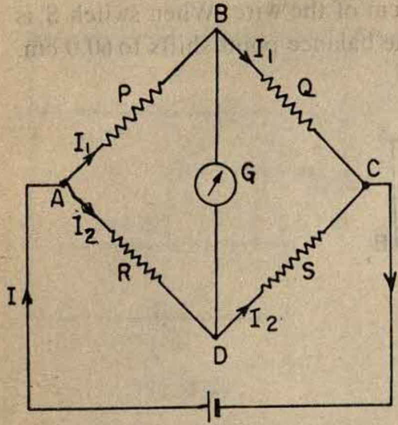


Fig. 3.41

Two of the known resistances, say P and Q, are fixed and the third resistance S is varied until the galvanometer shows zero deflection, i.e. no current flows in BD. When this happens, the bridge is said to be *balanced* and the resistances P, Q, R and S are related as

$$\frac{P}{Q} = \frac{R}{S}$$

or
$$R = \frac{PS}{Q} \tag{3.33}$$

Thus knowing the values of P, Q and S, the value of R is determined.

The balance condition (3.33) follows simply from Kirchhoff's rules. Suppose a current I enters junction A, where it divides into I_1 and I_2 as shown in the figure such that

$$I = I_1 + I_2$$

when the bridge is balanced, no current flows along BD. Hence the currents in arms BC and DC are I_1 and I_2 respectively. Now applying Kirchhoff's second rule to loops ABDA and BCDB we have

$$I_1P = I_2R$$

and
$$I_1Q = I_2S$$

Dividing the two we get the balance condition (3.33).

The Wheatstone bridge is a null method. Hence the internal resistance of the cell and the resistance of the galvanometer do not come into the picture; they do not affect the null point.

EXAMPLE 3.25 Figure 3.42 shows a metre bridge (which is nothing but a practical Wheatstone bridge) consisting of two resistances X and

Y together in parallel with a metre long constantan wire AC of a uniform cross-section. D is a movable contact that can slide along the wire AC. The resistors X, Y and resistances of segments AD and DC of the wire constitute the four arms of the bridge. The length of wire AC is 100 cm. X is a standard 4.00 Ω resistor and Y is a coil of wire. With Y immersed in melting ice, the movable contact D is moved along the wire until the galvanometer deflection is zero; the null point is found to be at a distance of 40.0 cm from point A. When coil Y is heated to 100°C, a 100 Ω resistor had to be connected in parallel with Y in order to keep the bridge balanced at the same point. Calculate the temperature coefficient of resistance of the coil.

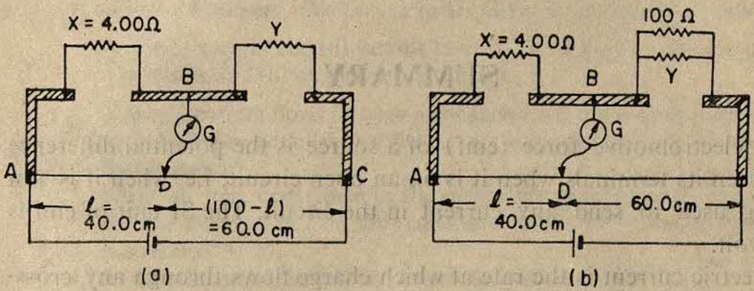


Fig. 3.42

Solution: Since the wire AC is of a uniform cross-section, the resistances of the two segments of the wire AD and DC are in the ratio of their lengths. If R_0 is the resistance of Y in melting ice (0°C), the balance condition of Wheatstone bridge gives

$$\frac{X}{R_0} = \frac{kl}{k(100 - l)} = \frac{l}{(100 - l)}$$

where k is the resistance per centimetre of wire AC. Now, $l = 40.0$ cm, and $X = 4.00\ \Omega$. Substituting these values, we get

$$R_0 = 6.00\ \Omega$$

Let R_t be the resistance of Y when heated to a temperature $t = 100$°C. When it is shunted with a 100 Ω resistor as shown in Fig. 3.42(b), the net resistance becomes

$$R' = \frac{100R_t}{R_t + 100}$$

Since the balance point remains unchanged, we have

$$\frac{X}{R'} = \frac{40.0}{60.0}$$

giving $$R' = 6.00\ \Omega$$

Thus $$6.00 = \frac{100R_t}{R_t + 100}$$

which gives $$R_t = 6.38\ \Omega$$

Temperature coefficient of resistance of the coil Y is

$$\beta = \frac{R_t - R_0}{R_0 t} = \frac{6.38 - 6.00}{6.00 \times 100} = 6.3 \times 10^{-4}\ \mathrm{K}^{-1}$$

SUMMARY

The electromotive force (emf) of a source is the potential difference between its terminals when it is on an open circuit, i.e. when it is not being used to send any current in the circuit. The SI unit of emf is the volt.

Electric current is the rate at which charge flows through any cross-section of a conductor; its SI unit is called the ampere.

Ohm's law states that the current flowing through a conductor is directly proportional to the potential difference across its ends, provided the physical conditions of the conductor (such as temperature) remain the same. The ratio of the potential difference across the ends of a conductor to the current flowing through it is called its resistance. The SI unit of resistance is called the ohm. The electrical resistivity of the material of a wire is the resistance of a unit length of the wire of a unit cross-sectional area. The reciprocal of electrical resistivity is called electrical conductivity. The resistance and resistivity of metals increase with increase in temperature.

The two Kirchhoff's rules are: (a) the algebraic sum of the currents at a junction is zero, and (b) the algebraic sum of the changes in potential across the sources of current and resistances in a closed loop is zero.

EXERCISES

A. Short-Answer Questions

1. What do you understand by electromotive force? State its SI unit.
2. State Ohm's law.

3. Define electrical resistivity. State its SI unit.
4. What is a superconductor?
5. Name two semiconductors.
6. What is the internal resistance of a cell due to?
7. Name the charge carriers in an electrovoltaic cell.
8. State the two Kirchhoff's rules.
9. Name the device used for measuring currents. How is it connected in a circuit?
10. Name the device used for measuring potential differences. How is it connected in a circuit?

B. Long Answer Questions

1. What constitutes electric current? Give a quantitative definition of: (i) steady current and (ii) varying current. Draw the current-time graph for: (i) a steady and (ii) an alternating current.
2. A steady current flows through a conductor of non-uniform cross-section. How will you establish that the current is the same for any cross-section of the conductor?
3. What do you understand by the terms drift velocity and relaxation time of the free electrons in a metallic conductor carrying a current? How are the two related?
4. A metallic wire of uniform cross-sectional area A has n free electrons per unit volume and a constant current I flows through it. Show that the magnitude v_d of the drift velocity of the free electrons is given by

$$v_d = \frac{I}{enA}$$

5. State Ohm's law for metallic conductors. Does the relation $V = IR$ hold for non-ohmic resistors? Explain the meaning of the term resistance in terms of the microscopic picture presented in this chapter.
6. Define the terms resistivity and conductivity of a conductor. What are their units?
7. Discuss the distribution of electron velocities in metals.
8. Discuss the variation of resistivity with temperature. Define the temperature coefficient of resistivity.
9. The temperature coefficient of resistance of a metal is $4.5 \times 10^{-4}\ \mathrm{K}^{-1}$. What is the meaning of this statement?
10. What is superconductivity? State two properties of superconductors. What is the most important application of superconductors?
11. N cells, each of emf $\mathcal{E}$ and internal resistance r are connected in series. The battery of cells is connected across an external resistance R. Find the current in the circuit. Discuss the special cases (i) $R \gg Nr$ and (ii) $R \ll Nr$.
12. N resistors of resistances $R_1, R_2, R_3, \ldots, R_N$ are connected in: (i) series (ii) parallel. Find the equivalent resistance of the combinations in (i) and (ii).
13. State Kirchhoff's rules and use them to find the potential differences across the resistors R_1, R_2, R_3 in the circuit shown in Fig. 3.25 of this chapter.

14. How can you convert a galvanometer into an ammeter? Describe with the aid of a concrete example.
15. With the help of an example, describe how a galvanometer can be converted into a voltmeter.
16. (a) Describe a potentiometer circuit arranged to compare the emf of a cell with that of a standard cell. (b) Explain the reasons for the following procedures in the experiment:

    (i) The positive terminal of the cell whose emf is required is connected to the same end of the potentiometer wire as the positive terminal of the driver cell.

    (ii) The protective resistor is removed before the final balance oint is determined.

    (iii) A rheostat is sometimes included in the potentiometer circuit with the driver cell but its resistance must not be too high.

17. In a potentiometer experiment, a balance length cannot be found. Suggest two possible reasons, with explanations and remedies.
18. Describe, with circuit diagrams, to show how a potentiometer is used to determine: (i) a low resistance and (ii) the internal resistance of a cell in terms of a standard resistor.
19. Describe clearly how a potentiometer is used to measure an emf of the order of 1 mV.
20. 'A potentiometer may be regarded as equivalent to a voltmeter.' Comment. Discuss the relative merits of using: (i) a potentiometer and (ii) a voltmeter to measure the emf of a cell.
21. Draw a Wheatstone bridge. Using Kirchhoff's rules, find the relationship between the resistances of the four arms of the bridge at null point.

C. Multiple-Choice Questions

Choose the correct answer or answers from the given alternatives.

1. Choose the only incorrect relationship from the following

   (a) Ampere = coulomb per second
   (b) Volt = coulomb per joule
   (c) Ohm = volt per ampere
   (d) Coulomb = joule per volt

2. The terminal voltage of a cell connected to an external resistance R is

   (a) always equal to its emf
   (b) always less than its emf
   (c) greater or less than its emf depending on the direction of the current through it.
   (d) greater or less than its emf depending on whether its internal resistance s less or greater than R.

3. A fresh flashlight cell of emf 1.5 V gives a current of 15 A when connected directly to an ammeter of resistance 0.04 Ω. The internal resistance of the cell is

   (a) 0.04 Ω (b) 0.06 Ω
   (c) 0.1 Ω (d) 1 Ω

4. The resistance of a straight conductor does not depend on its
   (a) temperature (b) length
   (c) material (d) shape of cross-section
5. A wire of resistance 4 Ω is stretched to twice its original length. What is the resistance of the wire now?
   (a) 1 Ω (b) 4 Ω
   (c) 8 Ω (d) 16 Ω
6. Each resistance in the network shown in Fig. 3.43 has a value of 6 Ω. The points P and A are connected by a copper wire of negligible resistance. The points Q and B are also connected by a copper wire of negligible resistance. What is the effective resistance between points P and Q?
   (a) 2 Ω (b) 3 Ω
   (c) 6 Ω (d) 18 Ω

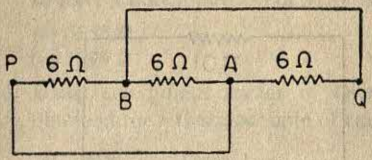


Fig. 3.43

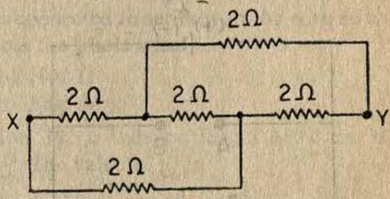


Fig. 3.44

7. The effective resistance between points X and Y in Fig. 3.44 is
   (a) 1 Ω (b) 2 Ω
   (c) 4 Ω (d) 6 Ω
8. Seven resistors, each of value 20 Ω, are connected to a 2 V battery as shown in Fig. 3.45. What is the reading of ammeter A?
   (a) 1/10 A (b) 3/10 A
   (c) 4/10 A (d) 7/10 A

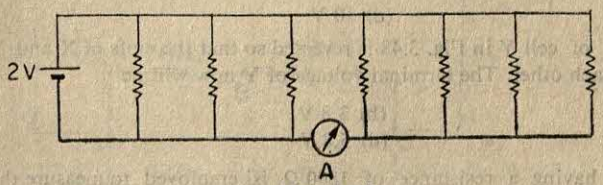


Fig. 3.45

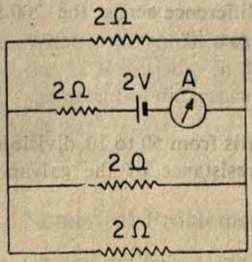


Fig. 3.46

9. The reading of ammeter A in the circuit shown in Fig. 3.46 is
   (a) 1/8 A
   (b) 3/8 A
   (c) 1/2 A
   (d) 2 A
10. A cell supplies a current of 0.9 A through a 2 Ω resistor and a current of 0.3 A through a 7 Ω resistor. The internal resistance of the cell is

(a) 0.5 Ω (b) 1.0 Ω
(c) 1.2 Ω (d) 2.0 Ω

11. What is the emf of the cell used in question 10 above?

(a) 1.5 V (b) 1.75 V
(c) 2.0 V (d) 2.25 V

12. Figure 3.47 shows a circuit with two cells X and Y in opposition to each other. Cell X has an emf of 6 V and internal resistance of 2 Ω and cell Y has an emf of 4 V and internal resistance of 8 Ω. What is the voltage between terminals A and B?

(a) 5.4 V (b) 5.6 V
(c) 5.8 V (d) 6.0 V

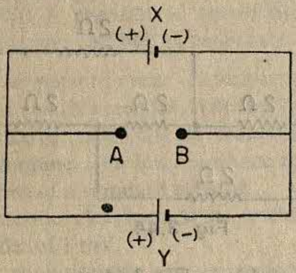


Fig. 3.47

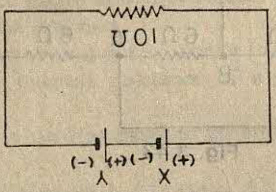


Fig. 3.48

13. Figure 3.48 shows two cells X and Y connected to a 10 Ω resistor. Cell X has an emf of 6 V and internal resistance of 2 Ω and cell Y has an emf of 4 V and internal resistance of 8 Ω. The terminal voltage of cell Y is

(a) zero (b) 2 V
(c) 4 V (d) 10 V

14. The polarity of cell Y in Fig. 3.48 is reversed so that the emfs of X and Y now oppose each other. The terminal voltage of Y now will be

(a) 2.8 V (b) 3.8 V
(c) 4.8 V (d) 5.8 V

15. A voltmeter having a resistance of 1800 Ω is employed to measure the potential difference across a 200 Ω resistor which is connected to the terminals of a dc power supply of emf 50 V and internal resistance 20 Ω. What is the percentage decrease in the potential difference across the 200 Ω resistor as a result of connecting the voltmeter across it?

(a) 1% (b) 5%
(c) 10% (d) 25%

16. The deflection in a moving coil galvanometer falls from 50 to 10 divisions when a shunt of 12 Ω is connected across it. The resistance of the galvanometer coil is

(a) 24 Ω (b) 36 Ω
(c) 48 Ω (d) 60 Ω

17. A galvanometer of resistance 10 Ω gives full-scale deflection for a current of 1.0 mA. What resistance must be connected in series with it so as to convert it into a voltmeter reading up to 2.5 V?

(a) 24.9 Ω (b) 249 Ω
(c) 2490 Ω (d) 24900 Ω

18. With what resistance should the galvanometer in Question 17 above be shunted so as to convert it into an ammeter that can measure currents up to 2.5 A?

(a) 4×10^{-1} Ω (b) 4×10^{-2} Ω
(c) 4×10^{-3} Ω (d) 4×10^{-4} Ω

19. The driver cell of a potentiometer has an emf of 2.0 V and negligible internal resistance. The potentiometer wire has a length 100 cm and resistance 5.0 Ω. What resistance must be connected in series with the wire so as to have a potential drop of 5 mV across the whole wire?

(a) 1985 Ω (b) 1990 Ω
(c) 1995 Ω (d) 2000 Ω

20. Using the potentiometer of Question 19, a balance length of 60 cm is obtained for a thermocouple of emf $\mathcal{E}$. The value of $\mathcal{E}$ is

(a) 3 mV (b) 4 mV
(c) 6 mV (d) 25/3 mV

21. In the circuit shown in Fig. 3.49, the potential difference between the points B and D is

(a) 1 V (b) 2 V
(c) 3 V (d) 4 V

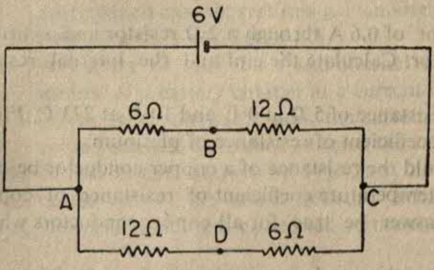


Fig. 3.49

22. What is the value of the resistance that should be connected in parallel with the 12 Ω resistor in branch ADC in the circuit shown in Fig. 3.49 so that the potential difference between points B and D becomes zero?

(a) 2 Ω (b) 4 Ω
(c) 6 Ω (d) 8 Ω

D. Numerical Problems

(In the following problems, take the electronic charge = 1.6×10^{-19} C and the electronic mass = 9.1×10^{-31} kg, whenever necessary).

1. 10^{19} electrons pass from a point A towards another point B in 1 s. What is the current in amperes? What is its direction?
2. A current of 2 A flows for 5 minutes in a conductor of resistance 10 Ω. (i) How many coulombs, and (ii) how many electrons pass through any cross-section of the conductor in this time?
3. What is the number of free electrons in a uniform wire of silver of area of cross-section 1.0×10^{-6} m^2 and length 1 m? Atomic mass of silver = 108, density of silver = 1.05×10^4 kg m^{-3}. Assume that there is one free electron in each atom of silver. The Avogadro number = 6.0×10^{23} mol^{-1}.
4. The wire considered in Problem 3 carries a current of 5 A. (i) What is the drift velocity of the free electrons in the wire? (ii) How long will the electrons take to drift from one end of the wire to the other?
5. A uniform copper wire of length 1 m and diameter 0.5 mm has a resistance of 0.08 Ω. Estimate τ, the average time between collisions of a free electron with the copper atom. Assume that there are 8×10^{28} free electrons in 1 m^3 of copper.
6. The average distance an electron travels between two collisions is called the mean free path. Assuming that the thermal speed of electrons in Problem 5 is about 10^8 cms^{-1}, calculate the mean free path.
7. A wire with a resistance of 5 Ω is drawn out so that its new length is twice its original length. What is the resistance of the longer wire?
8. A uniform wire has a resistance of 16 Ω. It is melted and made into a new uniform wire whose diameter is twice that of the original wire. What is the resistance of the new wire?
9. A cell has an emf of 1.5 V and an internal resistance of 1 Ω. It is connected to two resistance of 2 Ω and 3 Ω in series. Find: (i) the current flowing in the circuit and (ii) the potential difference across the ends of each resistance.
10. A cell supplies a current of 0.6 A through a 2 Ω resistor and a current of 0.2 A through a 7 Ω resistor. Calculate the emf and the internal resistance of the cell.
11. A platinum wire has a resistance of 5 Ω at 0°C and 10 Ω at 273°C. Find the value of the temperature coefficient of resistance of platinum.
12. At what temperature would the resistance of a copper conductor be double its resistance at 0°C? The temperature coefficient of resistance of copper is $4 \times 10^{-3}/^\circ$C. Will your answer be true for all copper conductors whatever their shape or size?
13. A galvanometer together with an unknown resistance in series is connected across two identical cell of emf 1.5 V. When the cells are connected in series, the galvanometer records a current of 1.0 A and when the cells are in parallel, the current is 0.6 A. Find the internal resistance of the cell.
14. A battery of emf 1.4 V and internal resistance 2 Ω is connected to a 100 Ω resistor through an ammeter whose resistance is 4/3 Ω. A voltmeter is also connected across the resistor. (i) If the ammeter reads 0.02 A, find the resistance of the voltmeter. (ii) If the voltmeter reads 1.10 V, what is the error in the reading?
15. What is the equivalent resistance between point A and B of the circuits shown in: (i) Fig. 3.50(a), (ii) Fig. 3.50(b), (iii) Fig. 3.50(c), and (iv) Fig. 3.50(d). Each resistor has a value of 10 Ω.

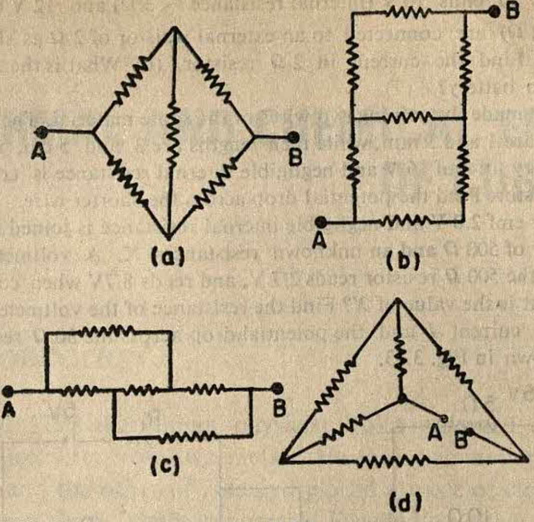


Fig. 3.50

16. A battery of 24 cells, each of emf 1.5 V and internal resistance 2 Ω, is to be connected so as to send the maximum current through a 12 Ω resistor. (a) How should the cells be connected? (b) Find (i) the current in each cell and (ii) the potential difference across the 12 Ω resistor.
17. A galvanometer having a coil resistance of 100 Ω gives full-scale deflection when a current of 1 mA is passed through it. Find the value of the resistance which can convert this galvanometer into an ammeter giving full-scale deflection for a current of 1 A.
18. A battery of emf 2.0 V and internal resistance 0.10 Ω is being charged by means of a battery charger at a current of 5 A. During the charging process (a) in what direction does the current flow inside the battery and (b) what is the potential difference between the two terminals of the battery?
19. In the circuit shown in Fig. 3.51, a voltmeter reads 30 V when it is connected across the 400 Ω resistor. What will be the reading of the same voltmeter when it is connected across the 300 Ω resistor ?

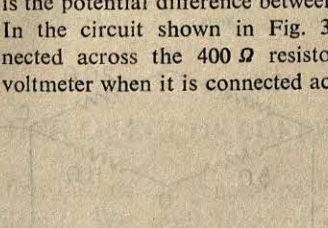


Fig. 3.51

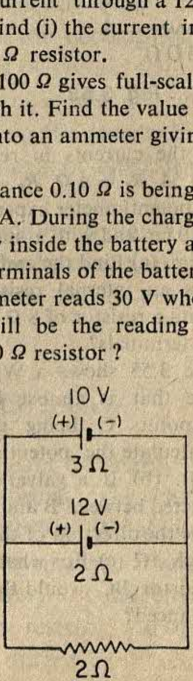


Fig. 3.52

20. Two batteries of emfs 10 V (internal resistance = 3 Ω) and 12 V (internal resistance = 2 Ω) are connected to an external resistor of 2 Ω as shown in Fig. 3.52 (a) Find the current in 2 Ω resistor. (b) What is the terminal voltage of each battery?
21. A resistor is made by joining two wires of the same material. The radii of the two wires are 1 and 3 mm, while their lengths are 3 and 5 cm, respectively. A battery of emf 16 V and negligible internal resistance is connected across the resistor. Find the potential drop across the shorter wire.
22. A battery of emf 2.0 V and negligible internal resistance is joined in series with a resistor of 500 Ω and an unknown resistances X. A voltmeter connected across the 500 Ω resistor reads 2/7 V, and reads 8/7V when connected across *X*. What is the value of *X*? Find the resistance of the voltmeter.
23. Calculate the current *I* and the potential drop across the 30 Ω resistor in the circuit shown in Fig. 3.53.

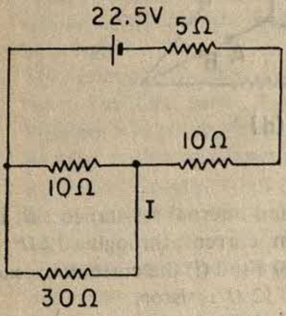


Fig. 3.53

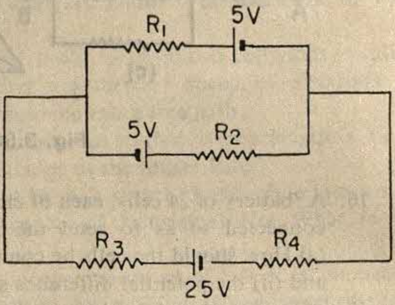


Fig. 3.54

24. Find the currents in resistors R_1, R_2, R_3 and R_4 in the circuit shown in Fig. 3.54. $R_1 = 4\ \Omega$, $R_2 = 12\ \Omega$, $R_3 = 10\ \Omega$ and $R_4 = 6\ \Omega$.
25. In a potentiometer experiment, the balance length with a cell A is found to be 75.0 cm and with a standard cell of emf 1.02 V, the balance length is 50.0 cm. (a) Find the emf of A. (b) What is the new balance length if cell A has an internal resistance of 2 Ω and a resistor of 8 Ω is joined to its terminals?
26. Figure 3.55 shows a Wheatstone bridge that is almost balanced, the point C being grounded. (a) Calculate the potentials at D and B. (b) If a galvanometer is connected between B and D, what will be the direction of the current through it? (c) For what value of the resistor BC would the bridge be balanced?

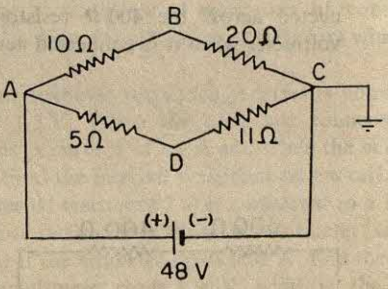


Fig. 3.55

4

THERMAL AND CHEMICAL EFFECTS OF CURRENTS

4.1 INTRODUCTION

In the year 1799 the Italian physicist Volta obtained a continuous supply of electricity from two metals. He took two metallic plates, one of copper and the other of zinc and placed a piece of cloth soaked in brine between them. This arrangement is called a voltaic cell. This is the beginning of a branch of physics called current electricity which deals with charges in motion.

The effects of electric current were discovered shortly after the discovery of the voltaic cell. It was found that water was decomposed into hydrogen and oxygen when connected to a series of voltaic cells. This was the earliest discovery of the chemical effect of an electric current. The heating effect of current was discovered soon afterwards, but the most important effect of current, namely, the magnetic effect was discovered some twenty years later.

In Chapter 3 we have studied the basic principles underlying the flow of charge. In this chapter we will discuss the heating and chemical effects of current. The magnetic effect of current will be dealt with in Chapter 5.

4.2 HEATING EFFECT OF CURRENT: JOULE'S LAW

We shall first discuss the general problem of transfer of energy in an electric circuit. We shall then consider the special case when the circuit contains a conductor having a finite resistance. We shall show that heat is developed in the conductor when a current flows through it.

Consider an electric circuit consisting of a battery B connected to a 'black box' [Fig. 4.1(a)]. This box may contain a resistor, a motor, an electrolyte, or anything else. A steady potential difference V is maintained between the terminals a and b of the black box and a steady current I flows in the circuit. Terminal a, which is connected to the

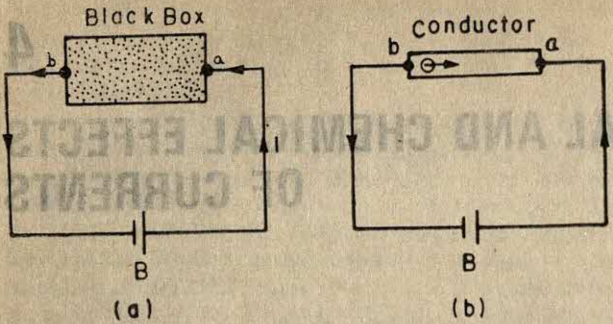


Fig. 4.1 *Energy transformation in an electrical circuit*

positive terminal of the battery, is at a higher potential than terminal b, which is connected to the negative terminal of the battery. Let us say that a charge q moves through the box from a to b in time t. Then the current I in the circuit is

$$I = \frac{q}{t} \tag{4.1}$$

In moving from a to b, the charge has to do work on whatever is contained in the box. This results in a decrease in its electrical potential energy, which can be obtained from the definition of V, the potential difference between points a and b. The potential difference between two points a and b in an electric field is defined as the amount of work done in moving a unit positive charge from a to b. If q units of charge move from a to b, the work done W by the charge is obviously q times V, i.e.

$$W = qV \tag{4.2}$$

This work represents the decrease in the electrical potential energy of the charge. The principle of conservation of energy tells us that this energy (like any other potential energy) is transformed (in the box) from the electrical potential energy to some other form of energy. To what other form the energy is transformed depends on what is contained in the box. From Eqs (4.1) and (4.2), we find that in time t the energy W transformed is given by

$$W = ItV \tag{4.3}$$

If the 'black box' contains a motor (such as a fan), the energy appears largely as mechanical work done by the motor. If the box contains an electrolyte, the energy appears largely as stored chemical energy (as in storage cells).

If the 'black box' contains a metallic conductor (i.e. a resistor), then the energy appears as heat in the resistor. This can be understood from the following example. Consider a stone of mass m falling through a height h. The decrease in its potential energy is mgh, where g is the acceleration due to gravity. If the stone falls in air (strictly, a vacuum), this energy is converted into the kinetic energy of the stone, resulting in its acceleration. But if the stone falls in water, it is progressively retarded as it falls in water, ultimately its speed becomes constant. The potential energy that is steadily being made available as the stone descends then appears in the form of heat in the surrounding water. The heat is developed due to the friction between the stone and water.

The motion of an electron through a conductor is somewhat like that of a stone through water. As the electron passes through the resistor, it falls, as it were, from a higher to a lower electrical potential. In the conductor [Fig. 4.1(b)], the electrons move from b to a, which is equivalent to a positive charge moving from a to b. The potential energy of an electron of charge $-e$ at b is $-eV_b$ and that at a is $+eV_a$. Since $V_a > V_b$, the electron loses potential energy as it moves from b to a. We have seen that electrons travel through a conductor with a constant drift velocity v_d and thus do not gain kinetic energy. The electrical potential energy that they lose is, therefore, not converted into kinetic energy. It is converted into heat energy in the conductor. This can be understood from the microscopic picture presented in Chapter 3. We have seen that the motion of electrons in a conductor, under the influence of an electric field, is hampered because the electrons suffer collisions with the atoms of which the conductor is made. It is for this reason that a conductor offers resistance to the motion of electrons (i.e.to the flow of the current). In between collisions, the electrons moving through the field acquire kinetic energy in addition to their own kinetic energy due to thermal motion. In the collisions, this energy is transferred to the atoms of the conductor. Consequently, the atoms begin to vibrate about their mean positions more and more violently. This increases the average kinetic energy of the atoms. The increase in the average kinetic energy of the atoms results in an increase in the temperature of the conductor. Thus, the conductor is heated by the flow of current through it.

Joule's Law

The heating of a conductor by the flow of an electric current through it is called *Joule heating*. This effect is thermodynamically irreversible.

We have seen that, in the case of a conductor, the electrical energy W given by Eq. (4.3) is converted into heat energy H, i.e.

$$H = ItV$$

where H is measured in energy units, i.e. joule.

For a conductor we have

$$V = IR$$

Combining these two equations, we have

$$H = I^2Rt \tag{4.4}$$

Equation (4.4) is called Joule's law which states that the amount of heat developed in a conductor by the passage of a steady current through it is proportional to: (i) the square of the current, (ii) the resistance of the conductor, and (iii) the time for which the current flows.

If I is measured in amperes, R in ohms and t in seconds, the heat energy H is the joules, i.e.

$$H = I^2Rt \text{ joules}$$

$$= \frac{I^2Rt}{J} \text{ cal}$$

where $J = 4.18 \text{ J cal}^{-1}$ is the mechanical equivalent of heat.

It may be mentioned that Eq. (4.3) applies to the conversion of electrical energy into any other form; but Eq. (4.4) applies only to the conversion of electrical energy into heat energy in the ohmic resistor. Equation (4.4) is a statement of Joule's law and is a particular way of writing the principle of conservation of energy for the special case in which electrical energy is converted into heat energy.

Electric Power

The rate at which electrical energy W is converted into other forms of energy is known as electric power P, which is given by

$$P = \frac{W}{t} = IV \tag{4.5}$$

i.e. Electric power = current × voltage.

In the SI system, the unit of electric power is ampere-volt. Since,

$$1 \text{ A} = 1 \text{ C s}^{-1}$$

and

$$1 \text{ V} = 1 \text{ J C}^{-1}$$

We have,

$$1 \text{ ampere-volt} = 1 \text{ A} \times 1 \text{ V}$$
$$= 1 \text{ J s}^{-1}$$

1 joule per second is called 1 *watt* (W). Thus

$$\textit{Watt} = \textit{ampere} \times \textit{volt}$$

Watt-hour Since 1 Watt = 1 J s^{-1}, it is clear that

$$1 \text{ watt-second} = 1 \text{ J}$$

or

$$1 \text{ watt-hour} = 3600 \text{ J}$$

Watt-second is a unit of work (or energy) and is equal to a joule. Watt-hour (= 3600 J) is a bigger unit of energy. The electric supply undertaking uses a much bigger unit for electrical energy consumed while calculating our electricity bill. This unit is called the kilowatt hour (kWh)

$$1 \text{ kilowatt hour} = 1000 \text{ watt-hour}$$
$$= 1000 \times 3600 \text{ joules}$$
$$= 3.6 \times 10^6 \text{ J}$$

The electric meter installed in our houses measures the electrical energy consumed in terms of kilowatt hours. Note that the watt (or kilowatt) is a unit of electrical power and watt-second (or watt-hour or kilowatt-hour) is a unit of electrical energy.

EXAMPLE 4.1 A cell of emf 1.5 V and internal resistance 0.10 Ω is connected across a resistor R. If the current in the circuit is 2.0 A, find: (a) the rate of chemical energy consumption in the cell, (b) the rate of energy dissipation inside the cell, (c) the rate of energy dissipation in the resistor, (d) the value of R, and (e) the power output of the source.

Solution: $V = 1.5$ V, $I = 2.0$ A, $r = 0.10\,\Omega$.

(a) Rate of chemical energy consumption in the cell $(P) = VI$

$$= 1.5 \times 2.0 = 3.0 \text{ W}$$

(b) Rate of energy dissipation in the cell $(P_C) = I^2 r$

$$= (2.0)^2 \times 0.1 = 0.4 \text{ W}$$

(c) Rate of energy dissipation in the resistor $R(P_R)$

$$= P - P_C = 3.0 - 0.4 = 2.6 \text{ W}$$

(d) Potential drop across $R = \dfrac{1.5 \times R}{R + 0.10}$

$\therefore$ Rate of energy dissipation in R = potential drop across $R \times$ current in R

$$= \frac{1.5 \times R \times 2.0}{R + 0.10}$$

which is $P_R = 2.6$ W. Hence

$$\frac{1.5 \times R \times 2.0}{R + 0.10} = 2.6$$

which gives $R = 0.65\,\Omega$

(e) Power output of the source = rate of dissipation of energy in the resistor

$$= 2.6 \text{ W}$$

EXAMPLE 4.2 An electrical cable having a resistance of $0.2\,\Omega$ delivers 10 kW at 220 V dc to a factory. Find the efficiency of the transmission.

Solution:

$$\text{Power loss in the cable} = I^2R = \left(\frac{P}{V}\right)^2 R$$

$$= \left(\frac{10{,}000}{220}\right)^2 \times 0.2$$

$$= 413 \text{ W} = 0.413 \text{ kW}$$

Now,

$$\text{Efficiency} = \frac{\text{power delivered by the cable to the factory}}{\text{power supplied to the cable}}$$

$$= \frac{10}{10 + 0.413} = 0.96 \quad \text{or} \quad 96\%.$$

EXAMPLE 4.3 A series battery of six lead accumulators, each of emf 2.0 V and internal resistance $0.50\,\Omega$ is charged by a 100 V dc supply. (a) What series resistance should be used in the charging circuit in order to limit the current to 8.0 A? (b) Using the required resistance, obtain: (i) the power supplied by the dc source, (ii) the power dissipated as heat and (iii) the energy stored in the battery in 15 min.

Solution:

(a) Total emf of the battery = $2.0 \times 6 = 12$ V

Total internal resistance of the battery $(r) = 0.50 \times 6 = 3.0\,\Omega$. While charging, the emfs of the dc supply and of the battery act in opposition since the current is sent in the battery from the negative to the positive terminal. Hence the effective emf $= 100 - 12 = 88$ V. If R is the required resistance, charging current $= 88/(R + 3) = 8.0$ A (Given) which gives $R = 8.0\,\Omega$.

(b) (i) Power supplied by the dc source $= 100 \text{ V} \times 8.0 \text{ A}$

$$= 800 \text{ W}$$

(ii) Power dissipated as heat in resistor $R = I^2(R + r)$

$$= (8.0)^2 \times (8.0 + 3.0)$$

$$= 704 \text{ W}$$

(iii) Energy stored in the battery in 15 min ($= 15 \times 60$ s)

$$= (800 - 704) \times (15 \times 60)$$

$$= 86400 \text{ J}$$

EXAMPLE 4.4 (a) A battery of emf $\mathcal{E}$ and internal resistance r is connected across a pure resistive device (such as an electric heater) of resistance R. Show that the power output of the device is maximum when $R = r$. Find the maximum power output. (b) What is the power output of the source if the battery is shorted? What is the power dissipation inside the battery in that case?

Solution: (a) The current in the circuit is

$$I = \frac{\mathcal{E}}{(R + r)}$$

$\therefore$ Power output of the device is given by

$$P = I^2R = \frac{\mathcal{E}^2R}{(R + r)^2} \tag{i}$$

For given values of $\mathcal{E}$ and r, power output P will be maximum if $dP/dR = 0$ and $d^2P/dR^2 < 0$. Differentiating (i) with respect to R we get (with $\mathcal{E}$ and r fixed)

$$\frac{dP}{dR} = \frac{\mathcal{E}^2}{(R + r)^2}\left\{1 - \frac{2R}{(R + r)}\right\} \tag{ii}$$

Now $dP/dR = 0$ if

$$1 - \frac{2R}{(R + r)} = 0$$

which gives $$R = r$$

Thus, P will be either maximum or minimum when $R = r$. To decide whether P is maximum at $R = r$, we find d^2P/dR^2 at $R = r$.

If its value is negative, P will be maximum. Differentiating (ii) we have

$$\frac{d^2P}{dR^2} = \frac{2\mathcal{E}^2}{(R+r)^3}\left\{\frac{3R}{(R+r)} - 2\right\}$$

$$\therefore \qquad \left(\frac{d^2P}{dR^2}\right)_{\text{at } R=r} = -\frac{\mathcal{E}^2}{8r^2}$$

which is negative. Hence P is maximum when $R = r$. Putting $R = r$ in Eq. (i) gives

$$(P)_{\max} = \frac{\mathcal{E}^2}{4r}$$

(b) When the battery is shorted, it cannot deliver any current in resistor R. Hence the power output of the device becomes zero. The current through the battery, in this case is $\mathcal{E}/r$ and the power dissipated inside the battery = (current)^2 × resistance = $(\mathcal{E}/r)^2 \times r = \mathcal{E}^2/r$.

EXAMPLE 4.5 (a) An electric motor runs on a dc source of emf $\mathcal{E}$ and internal resistance r. Show that the power output of the source is maximum when the current drawn by the motor is $\mathcal{E}/2r$. What is the maximum power output of the source. (b) Show that the power output of electric motor is maximum when the back emf is one-half the source emf provided the resistance of the windings of the motor is negligible. (c) Compare and contrast this situation with that in Example 4.4.

Solution: (a) When the motor is running (it is an active device), a back emf, say $\mathcal{E}'$, is developed in the circuit. If I is the current in the circuit, the power output of the source is given by

$$P = (\mathcal{E} - \mathcal{E}')I$$

But $\mathcal{E}' = Ir$. Hence

$$P = \mathcal{E}I - I^2r \qquad \text{(i)}$$

Differentiating (i) with respect to I we have

$$\frac{dP}{dI} = \mathcal{E} - 2Ir \qquad \text{(ii)}$$

P will be maximum or minimum if $dP/dI = 0$ i.e. if

$$I = \frac{\mathcal{E}}{2r}$$

Differentiating (ii) with respect to I gives

$$\frac{d^2P}{dI^2} = -2r$$

which is negative. Hence P is maximum when $I = \mathcal{E}/2r$. Setting $I = \mathcal{E}/2r$ in (i) we have

$$P_{\max} = \frac{\mathcal{E}^2}{4r}$$

(b) If the external resistance R is negligible,

$$I = \frac{\mathcal{E} - \mathcal{E}'}{r}$$

Since $R = 0$, Power output of the motor, P' = power output of the source

$$= (\mathcal{E} - \mathcal{E}')I$$

$$= (\mathcal{E} - \mathcal{E}')\frac{\mathcal{E}'}{r} \qquad (\because \ \mathcal{E}' = Ir)$$

or

$$P' = \frac{1}{r}(\mathcal{E}\mathcal{E}' - \mathcal{E}'^2)$$

Differentiating P' with respect to $\mathcal{E}'$ and setting the derivative equal to zero gives $\mathcal{E}' = \mathcal{E}/2$. It is easy to see that the second order derivative of P' with respect to $\mathcal{E}'$ is negative when $\mathcal{E}' = \mathcal{E}/2$. Hence P' is maximum when $\mathcal{E}' = \mathcal{E}/2$ provided $R = 0$.

Also

$$(P')_{\max} = \frac{1}{r}\left(\mathcal{E}\frac{\mathcal{E}}{2} - \frac{\mathcal{E}^2}{4}\right) = \frac{\mathcal{E}^2}{4r}$$

(c) The situation in this example is more general than that in Example 4.4 because a motor is not a passive resistor; it develops a back emf. The reason why $(P)_{\max}$ in Example 4.4 equals $(P')_{\max}$ in this example is that the external resistance R is neglected, i.e. the motor acts as a passive resistor. Hence the maximum power output in the two examples is the same.

EXAMPLE 4.6 A 64 V dc power supply is used to charge a battery of eight lead accumulators, each of emf 2.0 V and internal resistance of 1/8 Ω. The charging current also runs an electric motor connected in series with the battery. The resistance of the windings of the motor is 7.0 Ω and the steady current supply is 3.5 A. During 1 hour of charging, find: (a) the mechanical energy yielded by the motor and (b) the chemical energy stored in the battery.

Solution: The power supplied by the power supply is

$$P = 64 \text{ V} \times 3.5 \text{ A} = 224 \text{ W}$$

A part of this power is lost due to heat in the battery and in the windings of the motor. The remaining power is available for mechanical power yielded by the motor and the chemical power stored in the battery. Now

Rate of dissipation of heat energy in the battery

$$= (\text{current})^2 \times (\text{resistance of battery})$$

$$= (3.5\ \text{A})^2 \times \left(8 \times \frac{1}{8}\ \Omega\right)$$

$$= (3.5\ \text{A})^2 \times (1\ \Omega) = 12.25\ \text{W}$$

Rate of dissipation of heat energy in the windings of the motor

$$= (\text{current})^2 \times (\text{resistance of windings})$$

$$= (3.5\ \text{A})^2 \times (7.0\ \Omega) = 85.75\ \text{W}$$

∴ Total power lost due to heat dissipation

$$= 12.25 + 85.75 = 98\ \text{W}$$

Hence, the available power $= 224 - 98 = 126$ W. Now,

Chemical power stored in battery $=$ (voltage of battery) $\times$ (current)

$$= 16\ \text{V} \times 3.5\ \text{A} = 56\ \text{W}$$

∴ Mechanical power yielded by motor $= 126 - 56 = 70$ W

(a) Mechanical energy yielded by motor in 1 hour

$$= 70 \times 60 \times 60 = 252000\ \text{J}$$

(b) Chemical energy stored in the battery in 1 hour

$$= 56 \times 60 \times 60 = 201600\ \text{J}$$

EXAMPLE 4.7 A copper wire of cross-sectional area 0.50 mm^2 and length 10 cm is initially at 25°C and is thermally insulated from the surroundings. If a current of 10 A is set up in the wire: (a) Find the time in which the wire will start melting. Neglect the change in the resistance of the wire with temperature. (b) What will this time be if the length of the wire is doubled? Given, density of copper $= 9.0 \times 10^3$ kg m^{-3}, specific heat of copper $= 3.8 \times 10^{-1}$ J kg^{-1} K^{-1}, melting point of copper $=$ 1075°C and its resistivity $= 1.6 \times 10^{-8}\ \Omega$m.

Solution:

(a) Mass of copper wire (m) $=$ volume $\times$ density

$$= (0.5 \times 10^{-6}) \times 0.1 \times 9.0 \times 10^{-3}$$

$$= 4.5 \times 10^{-4}\ \text{kg}$$

Rise in temperature (ΔT) $= 1075 - 25 = 1050$°C

Heat energy required to melt the copper wire is

$$I^2Rt = m\times \Delta T\times C \tag{i}$$

where I = current = 10 A, R = resistance of wire, t = time required to melt the wire and C = specific heat of copper. Now

$$R = \frac{\rho L}{A} \tag{ii}$$

where ρ = resistivity, L = length and A = cross-sectional area of the wire. Using Eq. (ii) in Eq. (i) we have

$$t = \frac{m\Delta tCA}{I^2\rho L} \tag{iii}$$

Substituting for m, ΔT, C, A, ρ and L in (iii) and solving we get

$$t = 5.3\times 10^2 \text{ s}$$

(b) If the length (L) of the wire is doubled, its mass (m) is also doubled. It follows from Eq. (iii) that t remains the same, i.e. the wire will start melting at the same time.

4.3 CHEMICAL EFFECT OF CURRENT

The chemical effect of an electric current is one of the earliest effects of current known to man. One of the earliest industrial applications of electricity was the coating of a metal with a layer of another metal. This is called electroplating. Its purpose is to protect the metal from atmospheric corrosion and to give it a more attractive appearance. At one time, nickel plating was very popular; but nowadays chromium and platinum plating is used. The process involved in electroplating is called electrolysis.

Electrolysis

Almost immediately after the invention of the simple cell by Volta, now called the voltaic cell, two professors in England, Anthony Carlisle and William Nicholson, discovered a new phenomenon called electrolysis. They connected two platinum wires to a voltaic cell and dipped them in a vessel containing water to which a few drops of sulphuric acid had been added. They observed clouds of bubbles arising from each wire. They collected these gases in separate bottles. When the gases were tested it was found that one of them was hydrogen and the other oxygen.

It was already known that water consists of hydrogen and oxygen in a chemical combination. Carlisle and Nicholson showed, for the first time, that water could be decomposed into its elements by passing an electric current through it. This discovery was of profound importance because scientists thought that it was possible to discover the chemical composition of other substances by passing a current through them and thus decomposing them. This indeed was done a few years later by Humphry Davy and Michael Faraday.

Terms Used in Connection with Electrolysis

The process by which a substance is decomposed by the passage of an electric current is called *electrolysis*. The substance which conducts electricity and undergoes decomposition is called an *electrolyte*. The two wires (plates) which are connected to the battery are called *electrodes*. The electrode at which the current enters the electrolyte is called the *anode* and that by which the current leaves it is called the *cathode*. The vessel containing the electrolyte and the electrodes in which electrolysis is carried out is called an electrolytic cell or *voltameter*.

Water Voltameter

A simple form of a voltameter for the study of the electrolysis of water is illustrated in Fig. 4.2. It consists of two graduated tubes T_1 and T_2 provided with platinum electrodes P_1 and P_2. The tubes are connected

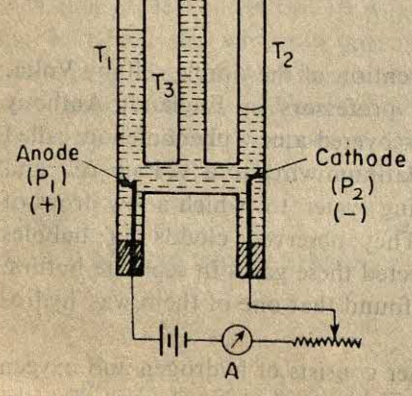


Fig. 4.2 *Water voltameter*

to a reservoir R as shown. Water with a few drops of sulphuric acid added to it (to make it conduct electricity) is poured into the reservoir till the tubes T_1 and T_2 are completely filled. The electrodes are connected to a battery. When current is passed through acidulated water, it is decomposed into oxygen and hydrogen. Hydrogen is liberated at the anode and oxygen at the cathode. When sufficient gas has been collected, it is observed that the volume of hydrogen is twice that of oxygen.

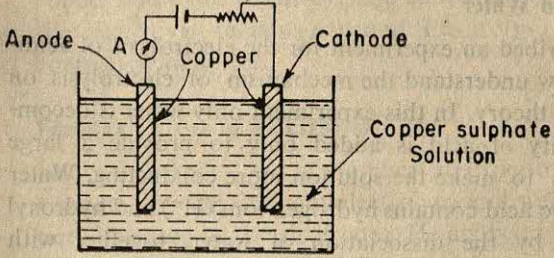


Fig. 4.3 *Copper voltameter*

Copper Voltameter

Figure 4.3 shows a copper voltameter which consists of a glass vessel containing copper sulphate solution as electrolyte and two copper plates as electrodes. When the electrodes are connected to a battery, copper sulphate solution decomposes into copper ions and sulphate ions. The copper ions move towards the negative electrode and get deposited there. The sulphate ions travel to the positive plate and combine with copper atoms of the plate forming copper sulphate. The copper sulphate so formed goes into solution and the concentration of the solution is preserved. Thus the cathode gains as much copper as the anode loses.

Ionic Theory of Electrolysis

We shall now explain why an electrolyte conducts electricity when it is dissolved in water. Take copper sulphate as an example. It consists of copper ions (Cu^{++}) and an equal number of sulphate ions (SO_4^{--}) arranged in a definite manner so that the whole is electrically neutral. A copper ion (Cu^{++}) is a copper atom which has lost two electrons and hence has a net positive charge ($+2e$). A sulphate ion is an SO_4 group that has gained the two electrons lost by the copper atom and hence has a negative charge ($-2e$). When copper sulphate is dissolved in water, the ions dissociate (or separate) and are able to move freely in the electrolytic solution. In fact, the word ion literally means 'wanderer'. When two electrodes are dipped in an electrolyte and connected to a battery, the positive ions (called cations) are attracted towards the negative electrode (i.e. the cathode) and the negative ions (called anions) are attracted towards the positive electrode (i.e. the anode). This migration of ions in both directions constitutes the flow of an electric current through an electrolyte.

Electrolysis of Acidified Water

We have already described an experiment for the electrolysis of acidified water. We can now understand the mechanism of electrolysis on the basis of the ionic theory. In this experiment only water is decomposed. A small quantity of acid is added only to provide a large number of extra ions to make the solution more conducting. Water acidified with sulphuric acid contains hydrogen ions (H^+) and hydroxyl ions (OH^-) produced by the dissociation of water, together with hydrogen ions (H^+) and sulphate ions (SO_4^{--}) produced by the dissociation of sulphuric acid.

When current is passed through the electrolyte, the H^+ ions travel to the cathode where they receive electrons and become neutral hydrogen atoms. The H atoms then combine in pairs to form molecules of hydrogen gas (H_2) which is liberated at the cathode. The action at the cathode may be represented by the following equations:

$$H^+ + e^- \rightarrow H$$

$$H + H \rightarrow H_2$$

While the H^+ ions are moving towards the cathode, the OH^- and SO_4^{--} ions move towards the anode. Here the pairs of OH^- groups combine to form water, H_2O, and oxygen O. The oxygen atoms combine in pairs to form molecules of oxygen gas (O_2) which is liberated at the anode. The SO_4^{--} ions remain in solution. The action at the anode may be represented by the following equations:

$$OH^- - e^- \rightarrow OH$$

$$OH + OH \rightarrow H_2O + O$$

$$O + O \rightarrow O_2$$

Electrolysis of Copper Sulphate Solution

Copper sulphate solution contains Cu^{++} and SO_4^{--} ions of copper sulphate together with H^+ and OH^- ions of water. When current is passed, Cu^{++} and H^+ ions travel to the cathode. The H^+ ions remain in solution. The Cu^{++} ions take two electrons of the cathode and become neutral Cu atoms, which are deposited on the cathode. After the current has passed for a sufficiently long time, the cathode becomes covered with a reddish layer of pure copper. At the anode, however, the SO_4^{--} ions combine with the copper ions of the anode, forming $CuSO_4$ which goes into solution, with the OH^- ions remaining in solution. Thus, as the electrolysis continues, the thickness of the copper layer on the cathode increases while the anode slowly loses copper and dissolves away. The concentration of copper sulphate in the electrolyte remains constant. Electrolysis stops when the anode is completely dissolved.

Faraday's Laws of Electrolysis

Michael Faraday did a number of experiments in which he passed currents through different electrolytes and weighed the substances that were liberated at the electrodes. He analysed the results of his experiments and discovered the following two laws.

First Law

The mass of the substance liberated in electolysis is proportional to the quantity of electricity (i.e. charge) passed through the electrolyte. If m is the mass of the substance liberated and q the total charge passed, then

$$m \propto q$$

or

$$m = zq = zIt \tag{4.6}$$

where I is the current passed in time t. The proportionality constant z is called the electrochemical equivalent (ece) of the substance. Its value depends on the nature of the substance. If $q = 1$ C, then $m = z$. Thus, the electrochemical equivalent of a substance is defined as that mass of it which is liberated during electrolysis when 1 C of electricity is passed through the electrolyte. The ece of a substance may be expressed in kilograms per coulomb.

$$\text{ece} = \frac{\text{Mass of the substance liberated (in kilograms)}}{\text{(Charge in coulombs)}}$$

$$= \frac{\text{Mass of the substance liberated (in kilograms)}}{\text{Current (in amperes)} \times \text{time (in seconds)}}$$

To determine the ece of copper, for example, the cathode in the experiment illustrated in Fig. 4.3 is weighed before the current is passed. A current of about 1 ampere (kept constant with the help of the rheostat) is passed for about half an hour. The cathode is weighed again. Knowing the mass of copper deposited, the current and the time for which the current is passed, we can calculate the value of the ece of copper.

Table 4.1 shows the value of the ece of a few substances.

Table 4.1 Electrochemical equivalents of substances

Element	Electochemical equivalent ($kg\ C^{-1}$)
Silver	1.118×10^{-6}
Copper	0.329×10^{-6}
Gold	0.681×10^{-6}
Oxygen	0.0829×10^{-6}
Hydrogen	0.0105×10^{-6}

Second Law

The masses of different substances liberated in electrolysis by the same quantity of electricity are proportional to their chemical equivalents. Thus, if m_1 and m_2 are the masses of substances of chemical equivalents e_1 and e_2, liberated when the same quantity of charge (e.g. when the same current is passed for the same time) is passed, then according to the second law

$$\frac{m_1}{m_2}=\frac{e_1}{e_2} \tag{4.7}$$

It may be mentioned that the chemical equivalent of an element is numerically equal to its atomic mass divided by its valency, i.e.

$$\text{Chemical equivalent}=\frac{\text{Atomic mass}}{\text{Valency}}$$

Thus, according to the second law, if the same current is passed through a solution of copper sulphate and acidified water for the same time, then for every one gram of hydrogen liberated in the electrolysis of water, 31.5 grams of copper will be liberated in the electrolysis of copper sulphate.

Relation between Electrochemical Equivalent and Chemical Equivalent

Suppose the same quantity of charge q is passed and m_1 and m_2 are the masses liberated of substances having electrochemical equivalents z_1 and z_2 respectively; then from the first law, we have

$$m_1=z_1q$$

and $$m_2 = z_2 q$$

or $$\frac{m_1}{m_2} = \frac{z_1}{z_2}$$

Combining this equation with Eq. (4.7) we have

$$\frac{e_1}{e_2} = \frac{z_1}{z_2} \quad \text{or} \quad \frac{e_1}{z_1} = \frac{e_2}{z_2}$$

which means that

$$\frac{e}{z} = \text{constant} = F \text{ (say)} \tag{4.8}$$

where F is called Faraday's constant.

To understand the meaning of F, let us use Eq. (4.6) in Eq. (4.8) to get

$$F = \frac{eq}{m}$$

If $m = e$, we have,

$$F = q$$

We know that the ece of copper is 0.000329 g C^{-1} and that its chemical equivalent is 31.5. Therefore, the amount of charge needed to liberate 31.5 g of copper is 31.5 g/0.000329 g C^{-1} = 96500 C. Thus, Faraday's constant represents 96500 C of charge.

The meaning of z, the chemical equivalent, may also be understood as follows: Consider a mole of an element liberated in electrolysis. Then the number of atoms of that element liberated is N, the Avogadro number. If the valency of the atom is p, then for each atom liberated, a charge pq_e must pass through the electrolyte. Here q_e is the magnitude of the electronic charge $= 1.6 \times 19^{-19}$ C. So, for one mole, the amount of charge passing through is Npq_e. If M is the mass of a mole, then in Eq. (4.6) $m = M$ and $It = q = Npq_e$. Hence Eq. (4.6) reads

$$M = zNpq_e$$

or $$z = \frac{1}{Nq_e} \cdot \frac{M}{p} \propto \frac{M}{p}$$

where the proportionality constant is $1/Nq_e$. The product Nq_e is called a Faraday (F), i.e. $F = Nq_e = q$. The quantity M/p (i.e. mass of a mole divided by valency) is called the chemical equivalent (symbol e), i.e.

$$e = \frac{M}{p}$$

Thus

$$z = \frac{e}{F} \quad \text{or} \quad F = \frac{e}{z}$$

which is Eq. (4.8). Thus Faraday's constant F is the amount of charge in coulombs required to liberate 1 mole of a substance; its value is 96500 C mol^{-1}.

Experimental Verification of Faraday's Laws

The first law can be verified with the experiment illustrated in Fig. 4.3. Pass a known current I (measured with ammeter A) for a certain length of time t. By adjusting the rheostat the current is kept constant. The cathode is weighed before and after passing the current and the mass m of copper deposited is determined. The experiment is repeated with double the current ($2I$) flowing for the same time t. It is observed the mass of copper deposited will be twice ($2m$) that in the first experiment. The experiment is repeated by passing the same current (I) as in first case but for twice the time ($2t$). It is observed that the mass deposited is again doubled. This verifies the first law.

The second law may be verified by taking three voltameters containing acidified water, copper sulphate solution and silver nitrate solution, respectively. They are connected in series, so that the same current passes through them. The electrolysis is carried out in each voltameter for the same time. The masses of hydrogen gas liberated, copper and silver deposited are determined. It is observed that for every 1 g of hydrogen liberated, 31.5 g of copper and 107.88 g of silver are deposited, thus verifying the second law.

Applications of Electrolysis

Electrolysis finds a variety of applications in industry.

Electroplating As mentioned earlier, electroplating is one of the earliest applications of electrolysis. The purpose is to give articles made of one metal a coating of another metal, in order to give them a more attractive appearance and to protect them against corrosion.

Purification of Metals Industrially, electrolysis plays a very important role in the refining of metals such as copper. Large plates of impure, crude copper serve as anodes placed in large vessels containing copper sulphate solution. The cathodes are thin sheets of pure copper. When a current is passed, pure copper ions from the anode are deposited on the cathode. Impurities at the anode are left behind and later removed.

Copper purified in this manner is called *electrolytic copper.* Zinc, aluminium, cadmium and other metals may be purified by electrolysis.

Chemical Analysis Sir Humphry Davy determined the chemical composition of many salts by electrolysing their solutions.

Manufacture of Chemicals Electrolysis is used in industry to produce pure sodium or potassium metal by electrolysis of sodium hydroxide with iron electrodes. Oxygen and hydrogen are manufactured by the electrolysis of water.

Printing Industry The types used in the printing of books are made using an electrolysis process. A wax impression of the type, for example the letter 'a', is made and the surface of wax is coated with graphite to make it conducting. This is made a cathode in a copper voltameter and a thin layer of copper is deposited on it by electrolysis. The copper is peeled off leaving the impression of the type on it, which can be used for printing. The discs of gramophone records are similarly coated with nickel.

EXAMPLE 4.8 For what time must a constant current of 2 A pass through a copper voltameter to deposit 1 g of copper? The electrochemical equivalent of copper is 0.000329 g C^{-1}.

Solution: We have

$$m = zIt$$

Here

$$m = 1\ \text{g}$$

$$z = 0.000329\ \text{g C}^{-1}$$

$$I = 2\ \text{A}$$

$$t = \frac{m}{zI} = \frac{1}{0.000329 \times 2}$$

$$= 1520\ \text{s}$$

$$= 25\ \text{min}\ 20\ \text{s}$$

EXAMPLE 4.9 An ammeter is suspected of giving inaccurate readings. In order to check the readings, it is connected in series with a silver voltameter and a steady current is passed for one hour. The ammeter indicates 0.54 A and 2.0124 g of silver is deposited. Is the ammeter reading correct? If not, what is the error? Thc electrochemical equivalent of silver of 0.001118 g C^{-1}.

Solution: The actual value of the current can be computed from

$$I = \frac{m}{zt}$$

Substituting the values of m, z and t, we have

$$I = \frac{2.0124}{0.001118 \times 3600}$$

$$= 0.50 \text{ A}$$

Hence, the current indicated by ammeter is higher than the actual current. The error is $(0.54 - 0.50) = +0.04$ A. It may be mentioned that this method of calibrating an ammeter, although somewhat laborious, is capable of good accuracy. In fact many years ago, the ampere used to be defined in terms of the mass of silver deposited per second in a silver voltameter.

EXAMPLE 4.10 A copper voltameter consists of two parallel-plate copper electrodes 5.0 cm apart, the surface area of each electrode being 1.0 m^2. The electrodes are immersed in a copper sulphate solution of resistivity $1.5 \times 10^{-2}\ \Omega$ m. Calculate the potential difference which must be established between the plates in order to deposit 0.48 kg of copper on the cathode in 1 hour. The value of Faraday's constant is 96500 C and the relative atomic mass of copper is 63.5.

Solution: The electrochemical equivalent of copper (because valency of copper = 2) is

$$z = \frac{M}{Fp} = \frac{63.5 \text{ g}}{(96500 \text{ C}) \times 2} = 3.29 \times 10^{-4} \text{ g C}^{-1}$$

$$= 3.29 \times 10^{-7} \text{ kg C}^{-1}$$

From $m = zIt$, we have

$$I = \frac{m}{zt} = \frac{0.48}{3.29 \times 10^{-7} \times 3600} = 4.05 \times 10^{2} \text{ A}$$

The resistance of the voltameter is

$$R = \frac{\rho l}{\mathbf{A}} = \frac{1.5 \times 10^{-2} \times 5.0 \times 10^{-2}}{(1.0)^2} = 7.5 \times 10^{-4}\ \Omega$$

From Ohm's law the potential difference across the plates is

$$V = IR = (4.05 \times 10^{2}) \times (7.5 \times 10^{-4})$$

$$= 3.03 \times 10^{-1} \text{ V}$$

$$\simeq 0.3 \text{ V.}$$

Note that we have used Ohm's law. It is found experimentally that Ohm's law is valid for voltameters with soluble electrodes (such as copper and silver voltameters). But for water and platinum voltameters, a 'back emf' must be taken into account before applying Ohm's law (see Example 4.11).

EXAMPLE 4.11 A steady potential difference of 1.62 V is maintained across two platinum electrodes placed in a solution of $CuCl_2$. At the end of 600 s, the mass of copper deposited on the cathode is measured to be 5.92 g. If the back emf of the voltameter is 1.34 V, find the resistance of the voltameter. The ece of copper $= 3.29 \times 10^{-7}$ kg C^{-1}.

Solution: $m = 5.92 \times 10^{-3}$ kg

Current through the voltameter is

$$I = \frac{m}{zt} = \frac{5.92 \times 10^{-3}}{3.29 \times 10^{-7} \times 600} = 29.99 \text{ A}$$

Effective potential difference $V = 1.62 - 1.34 = 0.28$ V. The resistance of the voltameter is

$$R = \frac{V}{I} = \frac{0.28}{29.99} = 9.34 \times 10^{-3}\ \Omega$$

EXAMPLE 4.12 The mass of hydrogen (H_2) evolved per unit coulomb of charge passed through a water voltameter is 1.05×10^{-8} kg C^{-1}. The corresponding figures for copper deposited in a copper voltameter and silver deposited in a silver voltameter are 3.29×10^{-7} kg C^{-1} and 1.12×10^{-6} kg C^{-1} respectively. The relative atomic masses of hydrogen, copper and silver are 1.008, 63.54 and 107.9 respectively. (a) What conclusion will you draw from these measurements? (b) Predict the amount of iron deposited per unit coulomb of charge passed through a voltameter with Fe electrodes in $FeCl_3$. The relative atomic mass of iron is 55.85.

Solution: (a) The valencies of hydrogen, copper and silver are 1, 2 and 1 respectively. The ratios of atomic mass and valency are 1.008 (for hydrogen), 63.54/2 = 31.77 (for copper) and 107.9 (for silver), i.e. their chemical equivalents are in the ratio of 1.008 : 31.77 : 107.9. From the measurements we notice that the masses deposited per unit coulomb are in the ratio of $1.05 \times 10^{-8} : 3.29 \times 10^{-7} : 1.12 \times 10^{-6}$ or 1.05 : 32.9 : 112. The two ratios are equal within the limits of experimental error. Therefore, we conclude that the masses deposited per unit coulomb are in proportion to the ratios of the chemical equivalents. This is Faraday's second law of electrolysis.

(b) The valency of Fe in $FeCl_3$ is 3. Therefore, the chemical equivalent of iron = 55.85/3 = 18.62. Using Faraday's second law, the mass of iron deposited per unit coulomb is

$$\frac{1.05\times10^{-8}\times18.62}{1.008}=1.94\times10^{-7}\text{ kg C}^{-1}$$

4.4 ELECTROCHEMICAL CELLS

Many chemical reactions are accompained by a release of energy. If this happens in an electrolyte, with one or more ionic species taking part in the reaction, the energy released will increase the potential energy of the ions. In such situations, the chemical reaction is the source of electrical energy. If the chemical reaction proceeds at a steady rate, the system can be used as a source of electrical power.

The energy released in a chemical reaction is of the order of 2×10^5 J per mole or $2\times10^5/6\times10^{23}$J per molecule. If the ionic charge on one of the reactants (or products) is 2 e (where $e = 1.6\times10^{-19}$C, the electronic charge), this energy will move the charge up by a potential V given by

$$2eV=\frac{2\times10^5}{6\times10^{23}}$$

or

$$V=\frac{2\times10^5}{2\times1.6\times10^{-19}\times6\times10^{23}}\simeq 1\text{ V}$$

Thus an electrochemical cell (i.e. a cell that converts chemical energy into electrical energy) has an emf of the order of 1 V. Known emfs are in the range 1 to 1.5 V, the highest being about 2.5 V.

Electrochemical cells are of two kinds, namely, *primary* cells and *secondary* cells. Secondary cells are also called *storage* cells or *accumulators*. Primary cells are those which produce an emf as a result of an irreversible chemical reaction, and hence cannot be recharged. Since the emf of a cell is derived from a chemical reaction, the total energy it can provide is limited by the amount of reactants which are gradually used up (as in the case of Voltaic, Daniell and Leclanche cells) unless the reactants are continuously supplied and products removed (as in the case of fuel cells). On the other hand, secondary cells produce an emf as a result of a reversible chemical reaction, and hence can be recharged.

Primary Cells

(a) *The Voltaic Cell* The simple voltaic cell consists of two plates, one of copper and the other of zinc, both dipped in dilute sulphuric acid contained in a glass vessel and kept apart from each other as shown in Fig. 4.4. The two plates are called the electrodes and dilute sulphuric acid is the electrolyte. As a result of a chemical action (discussed below), an emf of 1.08 V is developed between the plates. The copper plate is called the positive electrode (or terminal) and the zinc plate is the negative electrode.

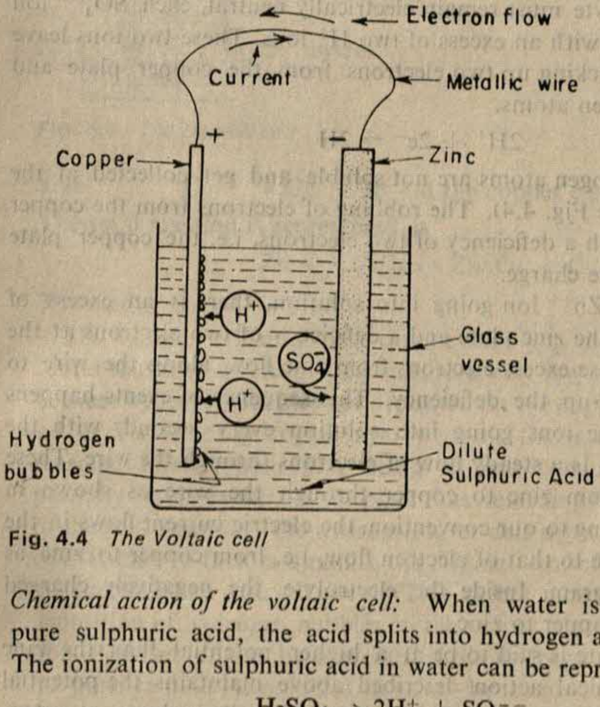


Fig. 4.4 *The Voltaic cell*

Chemical action of the voltaic cell: When water is added to dilute pure sulphuric acid, the acid splits into hydrogen and sulphate ions. The ionization of sulphuric acid in water can be represented as

$$H_2SO_4 \rightarrow 2H^+ + SO_4^{--}$$

The two hydrogen atoms lose one electron each and therefore acquire a positive charge. The sulphate ion takes the two electrons and acquires a negative charge.

The zinc metal is not soluble in water. However when it is placed in an acid solution, some of the zinc atoms ionize to become zinc ions which are soluble. The zinc atom ionizes by leaving electrons behind. All metals ionize like this. Each zinc atom which goes from the metal plate into the acid solution leaves two electrons behind on the plate. Having lost two electrons, each zinc atom becomes a doubly charged positive ion. The ionization of a zinc atom is represented by

$$Zn \rightarrow Zn^{++} + 2e^-$$

Let us now consider how the cell operates to produce a steady flow of electrons through a wire connecting the zinc and copper plates. A Zn^{++} ion goes into solution from the zinc plate, thereby leaving the plate with an excess of two electrons. Each Zn^{++} ion going into solution combines with SO_4^{--} ion to form zinc sulphate.

$$Zn^{++} + SO_4^{--} = ZnSO_4$$

Since the electrolyte must remain electrically neutral, each SO_4^{--} ion it loses, leaves it with an excess of two H^+ ions. These two ions leave the solution by picking up two electrons from the copper plate and becoming hydrogen atoms.

$$2H^+ + 2e^- = 2H$$

The neutral hydrogen atoms are not soluble and get collected at the copper plate (see Fig. 4.4). The robbing of electrons from the copper plate leaves it with a deficiency of two electrons, i.e. the copper plate acquires a positive charge.

Thus for each Zn^{++} ion going into solution, there is an excess of two electrons at the zinc plate and a deficiency of two electrons at the copper plate. These excess electrons from zinc flow along the wire to copper to make up the deficiency. This sequence of events happens for billions of zinc ions going into solution every second, with the result that there is a steady flow of electrons through the wire. These electrons flow from zinc to copper through the wire as shown in Fig. 4.4. According to our convention, the electric current flows in the direction opposite to that of electron flow, i.e. from copper to zinc as shown in the diagram. Inside the electrolyte, the negatively charged ions flow from copper to zinc.

The copper plate is said to be at a higher potential than the zinc plate. The chemical action described above maintains the potential difference between the plates. A simple voltaic cell produces a potential difference of 1.08 V between its terminals which is the emf of the cell.

(b) *The Daniell Cell* The Daniell cell consists of a zinc electrode immersed in dilute sulphuric acid and a copper electrode in copper sulphate solution, with a membrane (a procelain porous pot) through which the ions can pass from one solution to the other (Fig. 4.5). The zinc ions go into solution at the zinc (negative) electrode and copper ions are deposited at the copper (positive) electrode. The reactions are:

$$Zn \rightarrow Zn^{++} + 2e^- \text{ (at the zinc electrode)}$$

$$Zn^{++} + CuSO_4 \rightarrow ZnSO_4 + Cu^{++} \text{ (in solution)}$$

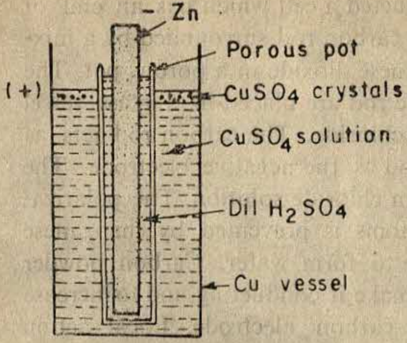


Fig. 4.5 *The Daniell cell*

and $Cu^{++} \rightarrow Cu - 2e^-$ (at the copper electrode)

The total reaction is represented as

$$Zn + CuSO_4 \rightarrow ZnSO_4 + Cu$$

The emf of the cell is 1.1 V.

The Voltaic or Daniell cells cannot supply a steady current. It is found that the current quickly falls to a very low value. This happens because hydrogen bubbles get deposited on the copper plate as explained earlier. The hydrogen layer weakens the main current for two reasons. Firstly, hydrogen, being a bad conductor of electricity, partially insulates the copper plate thus retarding the chemical action of the cell. Secondly, the hydrogen layer sets up a 'back emf' in the cell. These two factors diminish the original emf of the cell and hence the current in an external circuit. This defect which is caused by the deposition of hydrogen bubbles on the copper plate is called polarization. Polarization can be prevented if the hydrogen, before it reaches the positive plate, is oxidized to water by some oxidizing agent such as nitric acid, manganese dioxide, potassium dichromate, etc. These chemicals are called depolarizers. When a depolarizer is added to the acid, the hydrogen bubbles rapidly disappear due to their oxidation into water and the original emf of the cell is restored.

(c) *The Leclanché Cell* Any pair of dissimilar conducting materials can be put into any electrolyte and will become a cell which will convert chemical energy into electrical energy in a manner similar to the cells described above. Various, cells have been made using different electrodes and electrolytes. These cells vary widely in the amount of chemical energy released per coulomb of charge, which is just another way of saying that different cells have different emfs since the emf is the energy released in joules per coulomb of charge.

In 1865 Georges Leclanché constructed a cell which has an emf of about 1.5 V. The cell consists of a carbon rod surrounded by a mixture of powdered carbon and manganese dioxide in a porous pot. The porous pot and an amalgamated zinc rod are placed in a glass vessel containing a solution of ammonium chloride. The carbon rod acts as the positive electrode and the zinc rod as the negative electrode. The electrolyte (Fig. 4.6) is an ammonium chloride solution. The polarization of the cell by the hydrogen atoms is prevented by manganese dioxide which oxidizes the hydrogen to form water. Carbon powder is mixed with manganese dioxide to make it conducting and to increase the effective surface area of the carbon electrode. Local action at the zinc electrode is prevented by using amalgamated zinc.

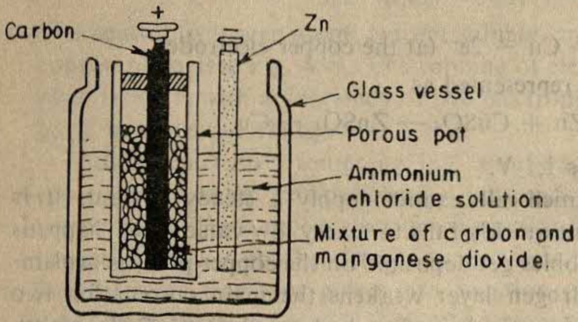

+
Carbon
−
Zn
Glass vessel
Porous pot
Ammonium chloride solution
Mixture of carbon and manganese dioxide.

Fig. 4.6 *The wet Leclanché cell*

The cell has one drawback in that it cannot maintain a steady emf of 1.5 V for a long time if a large current is drawn from the cell. The reason is that the depolarizing action of manganese dioxide is slow. Thus, if a large current is drawn from the cell (which implies that the number of hydrogen ions to be neutralized is large) the manganese dioxide cannot oxidize the hydrogen as fast as it is formed. Hence some hydrogen atoms get deposited on the carbon electrode and the cell is polarized, resulting in a fall in its emf. If the cell is allowed to rest for some time so that no further formation of hydrogen occurs but depolarizing action of manganese dioxide continues until all the hydrogen is oxidized to water, the cell regains its original emf and is ready for use again. It is for this reason that Leclanché cells are most suitable for use in the telephone, telegraph, electric bell, etc. where only intermittent current is drawn from the cell.

(d) *The Dry Cell* The cells described above are wet cells and are not easily portable because they use a solution as an electrolyte. The dry cell is a modified form of the wet Leclanché cell in which the ammonium chloride solution is replaced by a jelly consisting of starch, flour and ammonium chloride (Fig. 4.7). The positive terminal is a carbon rod surrounded by a compressed mixture of manganese dioxide and carbon. The carbon rod along with the surrounding depolarizing mixture is placed inside a zinc can and the intervening space is filled with the ammonium chloride jelly. The zinc can acts as the negative terminal. To prevent the jelly from drying up, the top of the cell is sealed by an insulating material. A cardboard washer is inserted at the bottom between the carbon rod and the zinc can so that the carbon rod does not come into contact with the zinc can at the bottom. The zinc can is surrounded by an insulating thick paper.

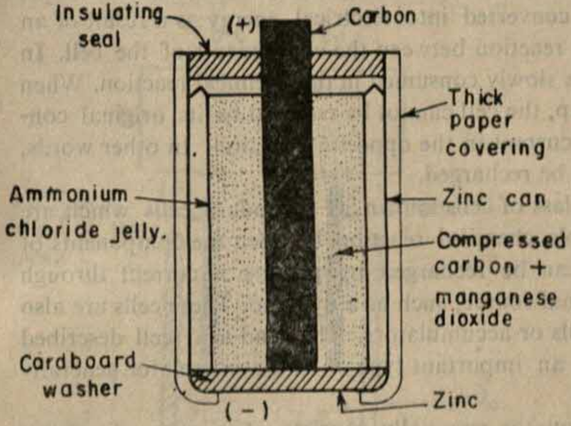


Fig. 4.7 *The dry Leclanché cell*

The emf of a dry cell is about 1.5 V. Dry cells are widely used in torches, transistor sets, tape recorders, etc. They are cheap and portable.

When an external circuit (e.g. a resistance wire) is connected across the cell, the zinc atoms in contact with the electrolyte ionize, losing two electrons per atom. The electrons flow through the wire and Zn^{++} ions pass into solution. Thus zinc is the negative terminal and the ionization at this terminal is represented by

$$Zn \rightarrow Zn^{++} + 2e^-$$

The ammonium ions of the electrolyte neutralize by combining with the electrons from the carbon electrode (which is the positive terminal) to which the electrons flow from the wire. Thus at the carbon electrode, we have

$$2\,NH_4^+ + 2e^- \rightarrow 2\,NH_3 + H_2$$

The hydrogen is neutralized by MnO_2 at the carbon electrode and in the electrolyte, the zinc ions (Z_n^{++}) combine with chlorine ions (Cl^-) to form zinc chloride. So the complete reaction is

$$Zn + 2NH_4Cl + MnO_2 \rightarrow ZnCl_2 + 2NH_3 + H_2O + Mn_2O_3 + \text{Energy}$$

This reaction releases energy which is stored in the cell as electrical energy producing an emf of 1.5 V.

Secondary Cell or Accumulator

The cells described above are known as primary cells. In these cells, chemical energy is converted into electrical energy as a result of an irreversible chemical reaction between the components of the cell. In these cells, the zinc is slowly consumed in the chemical reaction. When all the zinc is used up, the cell cannot be restored to its original condition by passing a current in the opposite direction. In other words, primary cells cannot be recharged.

There is another class of cells known as secondary cells which are based on a reversible chemical reaction between the components of the cell. These cells can be recharged by passing a current through them from an external source, such as a dynamo. These cells are also known as storage cells or accumulators. The lead-acid cell described in the following is an important type of lead accumulator generally used in laboratories.

A lead-acid accumulator essentially consists of a plate of spongy lead (which is the negative terminal) and another lead plate coated with lead oxide PbO_2 (which forms the positive terminal). The plates are immersed in sulphuric acid which is the electrolyte.

When a current is drawn from the cell (by connecting it to an external circuit), i.e. when the cell is *discharging*, the hydrogen ions drift to the positive plate and sulphate ions to the negative plate. As they give up their charges, they react with the plates and reduce the active materials of each plate to lead sulphate.

At the negative plate the reaction is

$$Pb + SO_4^{--} - 2e^- \rightarrow PbSO_4$$

The chemical action at the positive plate is given by

$$PbO_2 + H_2SO_4 + 2H^+ + 2e^- \rightarrow PbSO_4 + 2H_2O$$

But H_2SO_4 molecules do not exist in the solution; they are dissociated into $2H^+$ and SO_4^{--} ions. The above reaction may, therefore, be written as

$$PbO_2 + SO_4^{--} + 4H^+ + 2e^- \rightarrow PbSO_4 + 2H_2O$$

The lead sulphate formed in this reaction is spongy, and is more active than the usual hard form. In the above discharging reactions, water is formed and sulphuric acid is consumed and its density, therefore, falls.

Charging the Accumulator When the cell is discharged with use, it can be recharged by connecting it to a dc source of emf greater than that of the cell in such a way that the supply sends a current *I* through the cell in a direction opposite to that of the discharging current (see Fig. 4.8).

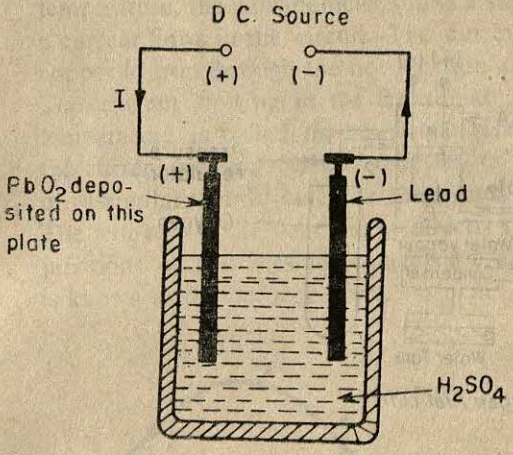


Fig. 4.8 *Charging a lead-acid accumulator*

During the charging process, the hydrogen ions are carried to the negative plate and the sulphate ions to the positive plate. The chemical reactions are as follows:

At the negative plate:

$$PbSO_4 + 2H^+ + 2e^- \rightarrow Pb + H_2SO_4$$

At the positive plate:

$$PbSO_4 + 2H_2O + SO_4^{--} - 2e^- \rightarrow PbO_2 + 2H_2SO_4$$

Thus we recover the active materials, namely, lead at the negative plate and lead oxide at the positive plate. During the charging process, water is consumed and sulphuric acid is formed. The acid, therefore, recovers its normal density and the cell is ready for use again.

The Fuel Cell

In fuel cells the chemical energy of fuels supplied from outside is converted into electrical energy. The main advantages of these cells are that there is no need to replace reactants (as in primary cells) or to recharge them (as in secondary cells). These cells can be put to large-scale use and are still in the process of development.

Figure 4.9 shows the essential parts of a hydrogen-oxygen fuel cell. It consists of an insulated vessel A fitted with electrodes B and C and containing potassium hydroxide which serves as the electrolyte. The electrodes are nickel plated plates with very fine pores that let the gases pass through. The surface coating of nickel acts as a catalyst for the chemical reactions at the electrodes. The fuel for the cell is hydrogen gas contained in a storage tank D. Hydrogen gas under regulated pressure is supplied to the electrode C where it combines with the hydroxide ions to form water. In this reaction, electrons are released. These electrons are received by electrode C which, therefore, becomes the negative terminal of the cell. Oxygen gas, from a storage tank E, is supplied under regulated pressure to the other electrode B. At this electrode, oxygen combines with water as well as electrons from this electrode to form hydroxide ions. Having lost electrons, this electrode becomes the positive terminal of the cell. The energy released in the chemical reactions at the electrodes is stored in the cell as electrical energy.

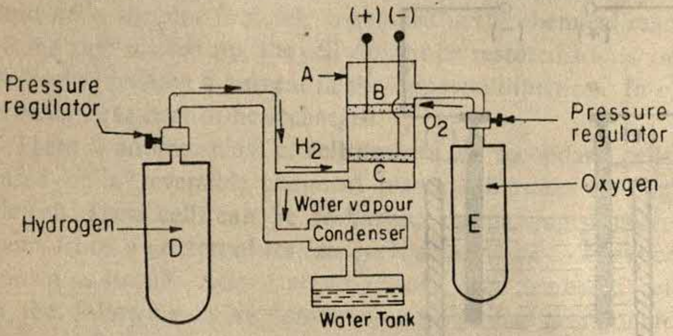


Fig. 4.9 *The hydrogen-oxygen fuel cell*

The chief merit of the fuel cells is that they are very efficient; the chemical energy can be completely converted into electrical energy. The main reasons why these cells have not yet been fully developed for large-scale use are the prohibitive cost of materials and the lack of long-term reliability.

4.5 THERMOELECTRICITY

Seebeck Effect

In 1826, Seebeck discovered that if a circuit is composed of two dissimilar metals and the junctions of these metals are maintained at different temperatures, an emf is set up in the circuit. Consequently a current flows in the circuit. Figure 4.10 shows two pieces of wire, one of copper and the other of iron, joined together at their ends at A and B. A sensitive galvanometer G is included in the circuit as shown. When one junction, say B, is heated and the other kept at a constant temperature, the galvanometer shows a small deflection indicating that a current flows in the circuit. The current so produced flows from copper to iron through the hot junction as indicated by the arrows. The current flowing in the circuit as a result of the difference in temperature is called *thermoelectric current* and the emf that derives this current is called *thermoelectric emf* or simply *thermo emf.* The pair of dissimilar metals joined at their ends constitutes a *thermocouple.* The conversion of heat energy directly into electric energy when the junctions of two dissimilar metals are kept at different temperatures is known as the *Seebeck effect.*

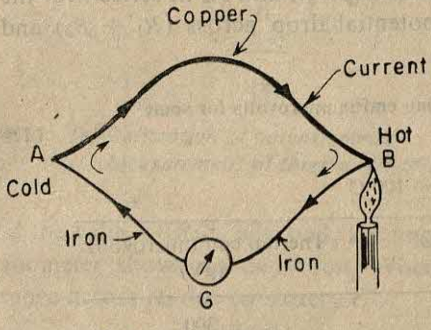


Fig. 4.10 *Seebeck effect*

In his original experiment, Seebeck used a thermocouple made of copper and bismuth. He could detect the current even when he held one junction in his hand. Later, he noticed a similar behaviour in many other pairs of metals. He found that the value of the thermo emf was different for different pairs of metals constituting the thermocouple when the same temperature difference was maintained between their junctions. He arranged the metals in a series such that when any two of them formed a couple, the current flowed at the hot junction from the metal occurring earlier in the series to the one occurring later in the series. Some of the substances in the series are:

Bi, Ni, Co, Pt, Cu, Mn, Pb, Au, Ag, Zn, Cd, Fe, Sb

The more separated the metals are in the series, the greater is the value of the thermo emf generated. Thus the emf developed in a Bi-Sb thermocouple is higher than that in a Cu-Zn thermocouple.

The sign and magnitude of the thermo emf depends on the two metals involved. The effect is reversible, i.e. if the hot and cold junctions are interchanged, the emf changes sign; its magnitude remaining unchanged.

Measurement of Thermo EMF

The thermo emf is measured by a potentiometer described in Chapter 3. The ordinary potentiometer cannot be used for measuring thermo emf because its value is very small, usually of the order of a few millivolts (see Table 4.2). Hence the balancing length of the potentiometer is very small and cannot be measured with any accuracy. Figure 4.11 illustrates a modified form of potentiometer used for measuring very small emf's. To obtain a reasonably large value of l, the potential difference across AB must be of the order of a few millivolts. This is achieved by connecting high resistances R_1 and R_2 in series with the wire AB, thus allowing a large potential drop across $(R_1 + R_2)$ and the rest across the wire AB.

Table 4.2 Magnitude of thermo emf in microvolts for some thermocouples

Temperature of cold junction = 0 °C
Temperature of hot junction = 100 °C

Thermocouple	Thermo emf (microvolt) (μV)
Lead-copper	300
Lead-Platinum	400
Iron-lead	1400
Copper-Constantan	4000
Iron-Constantan	5400
Bismuth-lead	6900

The first step is the determination of resistance r of wire AB. The next step is the calibration (i.e. determination of k) of the potentiometer. This is done by using a standard cadmium cell of known emf $\mathcal{E}_0 = 1.0184$ V (Fig. 4.11a). Resistances are taken out of R_1 and R_2 so that $(R_1 + R_2) \simeq 4000\,\Omega$. The jockey J is placed in contact with

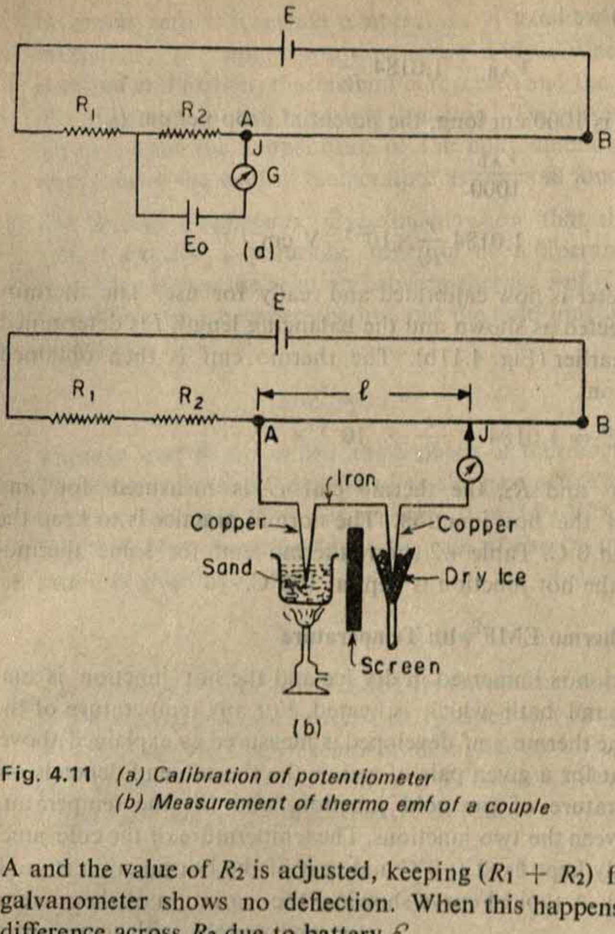


Fig. 4.11 *(a) Calibration of potentiometer*
(b) Measurement of thermo emf of a couple

A and the value of R_2 is adjusted, keeping $(R_1 + R_2)$ fixed, until the galvanometer shows no deflection. When this happens, the potential difference across R_2 due to battery $\mathcal{E}$

$$= \mathcal{E}_0, \text{ the emf of the standard cell}$$
$$= 1.0184 \text{ V}$$

Since $V_{R_2} = \dfrac{VR_2}{R_1 + R_2 + r}$, we have

$$\frac{V}{R_1 + R_2 + r} = \frac{1.0184}{R_2} \tag{4.9}$$

Now the potential V_{AB} across the wire AB is given by

$$V_{AB} = \frac{Vr}{R_1 + R_2 + r}$$

Using Eq. (4.9) we have

$$V_{AB} = 1.0184 \frac{r}{R_2}$$

If the wire AB is 1000 cm long, the potential drop per cm (k)

$$k = \frac{V_{AB}}{1000}$$

$$= 1.0184 \frac{r}{R_2} \times 10^{-3} \ \text{V cm}^{-1}$$

The potentiometer is now calibrated and ready for use. The thermocouple is connected as shown and the balancing length l is determined as mentioned earlier (Fig. 4.11b). The thermo emf is then obtained from the relation

$$\mathcal{E} = 1.0184 \times \frac{r}{R_2} \times 10^{-3} \times l$$

Knowing r, l and R_2, the thermo emf $\mathcal{E}$ is measured for any temperature of the hot junction. The normal practice is to keep the cold junction at 0°C. Table 4.2 shows thermo emf for some thermocouples when the hot junction is kept at 100°C.

Variation of Thermo EMF with Temperature

The cold junction is immersed in dry ice and the hot junction is embedded in a sand bath which is heated. For any temperature of the hot junction the thermo emf developed is measured as explained above. It is found that for a given pair of metals the thermo emf depends on: (i) the temperature of the cold junction, and (ii) the temperature difference between the two junctions. The temperature of the cold junction is normally kept fixed at 0°C and that of the hot junction is gradually increased. A typical curve showing the variation of thermo emf for any couple with temperature is shown in Fig. 4.12. The thermo emf increases as the temperature of the hot junction is raised, reaching a maximum value at a temperature t_N after which it begins to decrease. The temperature t_N is called the neutral temperature of the couple and is different for different couples. For a copper-iron couple $t_N \simeq 270$°C. On further raising the temperature of the hot junction, thermo emf decreases,

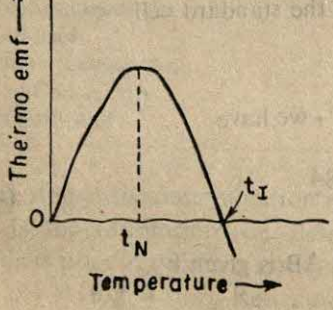


Fig. 4.12 *Variation of thermo emf with temperature*

becoming zero at a certain temperature t_I called the temperature of inversion. For copper-iron couple t_I is 540°C. When the temperature is raised still further, thermo emf is reversed and the current then flows in the couple in the opposite direction. The inversion temperature depends upon the temperature of the cold junction. It is always as much above the neutral temperature as the cold junction is below it.

The Seebeck Coefficients Experiments show that the thermo emf is almost exactly a parabolic function of temperature, as shown in Fig. 4.12. Hence the relation between thermo emf $\mathcal{E}$ and the temperature difference t between the hot and the cold junctions can be represented by

$$\mathcal{E} = \alpha t + \frac{1}{2}\beta t^2 \tag{4.10}$$

where α and β are called the Seebeck or thermoelectric coefficients. The values of α and β depend upon the pair of the metals constituting the thermocouple. For copper-iron, $\alpha = 14\ \mu\text{V/°C}$ and $\beta = -0.04\ \mu\text{V/(°C)}^2$. Now from Eq. (4.10) the rate of change of emf with temperature is given by

$$S = \frac{d\mathcal{E}}{dt} = \alpha + \beta t \tag{4.11}$$

The quantity $S = \mathrm{d}\mathcal{E}/\mathrm{d}t$ is called the *thermoelectric power* of the couple at the particular temperature; it is expressed in volts per degree Celsius. S is a linear function of temperature as shown in Fig. 4.13. The intercept of the line is α and the slope is β. Table 4.3 shows the values of α and β for some materials relative to lead.

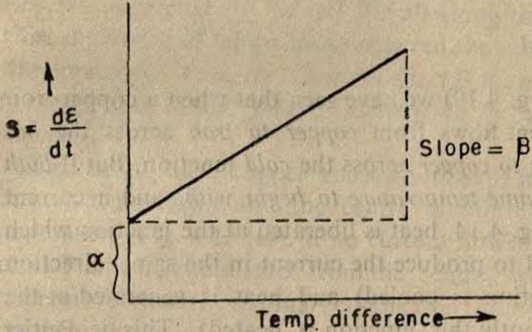


Fig. 4.13 *Thermoelectric power line of a thermocouple*

Peltier Effect

The converse of Seebeck's effect has also been observed. In 1834, a French scientist Jean Charles Athanase Peltier, discovered that when

Table 4.3 Values of α and β of some substances relative to lead

Substance	α (μV/°C)	β μV/(°C)^2
Aluminium	−0.76	+0.0039
Antimony	+35.6	+0.145
Bismuth (commercial)	−43.7	−0.465
Copper	+2.76	+0.012
Iron	+16.65	−0.030
Platinum	−3.04	−0.033
Silver	+3.44	+0.008
Constantan	−33.10	+0.080
Gold	+2.80	+0.010

current is passed across the junction of two different metals, heat is either evolved (generated) or absorbed (liberated) (i.e. the junction is either heated or cooled) depending upon the direction of the current. The quantity of heat evolved or absorbed depends upon the temperatures of the junctions and the materials of the conductors (Fig. 4.14). This discovery led to the explanation of the Seeback effect.

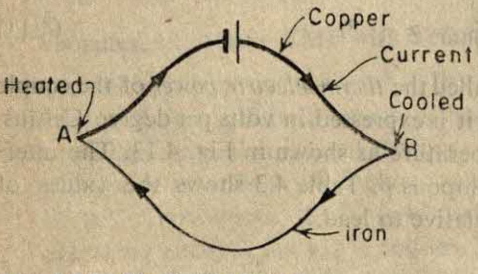


Fig. 4.14 *The Peltier effect*

In Seebeck's effect (Fig. 4.10) we have seen that when a copper-iron junction is heated, current flows from *copper to iron* across the *hot* junction and from *iron to copper* across the *cold* junction. But if *both the junctions are at the same temperature to begin with,* and a current is sent, as shown in Fig. 4.14, heat is liberated at the junction which should have been heated to produce the current in the same direction (consequently this junction is cooled) and heat is generated at the other junction (consequently this junction is heated). This is Peltier effect. If the Seebeck emf (i.e. thermo emf) is from copper to iron at the hot junction, an external emf applied in this direction will produce cooling at this junction and heating at the other. Thus the junction at which the current enters from copper to iron is cooled and that at which the current enters from iron to copper is heated (Fig. 4.14).

The Peltier Coefficient The amount of heat energy (in joules) evolved or absorbed at a function of two different metals when 1 coulomb of charge flows across the junction is called the Peltier coefficient. It is represented by the symbol π which is expressed in joules per coulomb or volt and its value depends upon the material of the two metals and the absolute temperature of the junction.

If the two junctions are at same temperature, their Peltier coefficients are equal. If one junction is at a higher absolute temperature T_2 than the other junction at absolute temperature T_1, their Peltier coefficients π_2 and π_1 will be different. If a charge q coulombs is passed through the circuit, the external work done $= q\mathcal{E}$ where $\mathcal{E}$ is the total emf in the circuit. The heat absorbed at the junction $= (\pi_2 - \pi_1) \times q$ where $T_2 > T_1$. This produces a net emf across the junction given by $\mathcal{E} = \pi_2 - \pi_1$. This is the Peltier emf.

Note: The Peltier effect is very different from the Joule heating effect. The Joule effect, which is always a heating effect, is proportional to the square of the current (I^2) and is, therefore, an irreversible effect. The Peltier effect, on the other hand, is directly proportional to current (I) and is reversible i.e. heat can be evolved or absorbed depending on the direction of the current. Hence the usual formula I^2R = power dissipated is not applicable to this situation.

Thomson Effect

The famous British physicist William Thomson (later Lord Kelvin) suggested that when a current flows in a thermocouple circuit, the absorption or evolution of energy does not occur merely at the junctions (as suggested by Peltier) but all along the conductors by virtue of the difference of temperature between the ends of each element of a thermocouple. Thus *the Thomson effect is the absorption (or evolution) of energy due to the flow of current in an unequally heated single conductor.*

Consider a thick copper rod with its ends maintained at the same constant temperature and the centre maintained at a much higher temperature [Fig. 4.15 (a)]. If no current flows in this conductor, the points P and Q, equidistant from the centre, would be at the same temperature due to thermal conduction alone. If now a current is sent through the rod in the direction shown in the figure, it is observed that the temperature at P is less than that at Q. This means that heat energy has been transferred from P to Q (in the direction of the current). This is called the *positive Thomson effect.* A similar effect is observed in Cd, Zn, Ag and Sb.

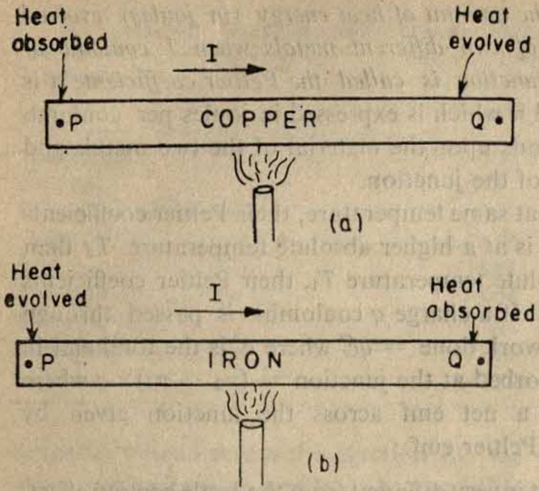


Fig. 4.15 *Thomson effect in: (a) copper and (b) iron*

Now consider an iron rod under similar conditions [Fig. 4.15 (b)]. When a current is sent in the iron rod in the direction from P to Q, the point P becomes hotter than point Q, i.e. the heat energy is transferred from Q to P (in a direction opposite to that of the current). This is called the *negative Thomson effect*. A similar effect is also observed in Pt, Bi, Co, Ni and Hg. If the direction of the current in either of the above cases is reversed, the Thomson effect is also reversed.

In lead, the Thomson effect is zero. It is for this reason that the thermoelectric behaviour of metals is studied by taking lead as the second element as in Table 4.3.

The Thomson Coefficient We have seen that in all conductors, except lead, energy is absorbed or evolved all along an unequally heated conductor when a current is made to flow through it. The difference in the energy absorbed or evolved all along the two metals, which form a thermocouple, gives rise to an emf round the thermocouple due to Thomson effect. The Thomson emf for a given metal A is given by $(\sigma_A)\,dT$ where dT is the difference in temperature between adjacent points on the metal. The symbol σ_A denotes the Thomson coefficient of A.

The Thomson coefficient (σ) *is defined as the energy absorbed or evolved when 1A current flows for 1s between two points of a conductor which differ in temperature by 1°C (or 1 K).* Thus if 1C of charge is carried from a point at absolute temperature T in a conductor to another point at temperature $(T + dT)$, the energy absorbed or evolved $= \sigma\, dT$. The Thomson coefficient (σ) is also defined as the potential difference set up between two points per unit difference of temperature between them i.e.

$$\sigma = \frac{\mathrm{d}\mathcal{E}}{\mathrm{d}T} \tag{4.12}$$

Relationship between the Seebeck, Peltier and Thomson Coefficients

Thomson showed that the Seebeck coefficient S (also called thermoelectric power), the Peltier coefficient π and the Thomson coefficient σ are related as follows:

$$S = \frac{\mathrm{d}\mathcal{E}}{\mathrm{d}T} = \frac{\pi}{T} \tag{4.13}$$

and

$$\sigma = -\,T\left(\frac{\mathrm{d}^2\mathcal{E}}{\mathrm{d}T^2}\right) = -\,T\left(\frac{\mathrm{d}S}{\mathrm{d}T}\right) \tag{4.14}$$

For a derivation of these relations, refer to the *supplement* at the end of this chapter.

The Origin of Thermoelectricity

We can understand thermoelectricity from our knowledge of the atomic structure of conductors. We know that every conductor has a large number of free electrons which wander about in the body of the conductor just like the molecules of a gas. The electrons have speeds up to a rather high value $\sim 10^6\ \mathrm{ms}^{-1}$. As the temperature is increased, the kinetic energy of the electrons increases. Thus if one end of a metal is hotter than the other, the electrons from the hotter end, being more energetic, diffuse faster than the electrons from the colder end. Thus there is a net electron flow which sets up an emf which counteracts and finally stops the flow of electrons. This is the Thomson effect.

The density (i.e. number/volume) of the electrons near the interface (end) of a conductor is, in general, different for different conductors since it depends on the temperature at the interface and the material of the conductor. Hence, when two dissimilar metals are brought into contact, the electrons tend to 'diffuse' (just like the molecules of two gases) from one conductor into the other. The rates of diffusion are different, since electron densities in the two metals are different. The metal that has undergone a net loss of electrons becomes positively charged while the one that has gained electrons becomes negatively charged. As the charges are built up, a contact potential difference is established across the interface (boundary) between two metals. When the temperatures of the two junctions are equal, there is no net emf in the circuit and consequently no current flows. When, however, one junction is heated, the contact potential at this junction becomes more than that at the cold junction; the difference between the two is the resultant emf which drives the current in the circuit. This is the Seebeck effect.

The contact potential difference $\mathcal{E}_{AB}$ is defined as the energy gained by an electron crossing junction AB from metal A to B (i.e. the interface between A and B) divided by the electron charge. So the work done W_{AB} to move an electron from metal A to metal B across the junction is

$$W_{AB} = W_A - W_B = e\mathcal{E}_{AB}$$

Where W_A and W_B are the work functions of A and B respectively. This net work done is stored in the system as electrical energy in the form of an emf $\mathcal{E}_{AB}$. Since, for any pair of conductors, $W_{AB} \ll e$, the value of $\mathcal{E}_{AB}$ is small; of the order of a few millivolts.

Applications of Thermoelectricity

The important applications of thermoelectricity are: (a) measurement of temperature, (b) detection of thermal radiations and (c) in devices such as refrigerators and generators.

(a) *Thermoelectric Thermometer* The thermo emf of a couple is given by

$$\mathcal{E} = \alpha t + \frac{1}{2}\beta t^2$$

The constants α and β are determined by measuring the thermo emf at two known standard temperatures, e.g. the melting points of certain solids. Any unknown temperature can then be computed from the above equation by measuring the corresponding emf produced at that temperature (see Example 4.13).

In a thermoelectric thermometer, the wires forming the couple are fused at the hot junction which is enclosed in a porcelain tube (Fig. 4.16). The wires to the cold junction are insulated and soldered at the end which is immersed in dry ice. A galvanometer is calibrated to directly indicate the temperature of the hot junction. As stated earlier, a copper-constantan thermometer is suitable for measuring temperatures of up to about 300°C. Thermocouples of platinum and platinum-rhodium alloy are used for measuring temperatures of up to

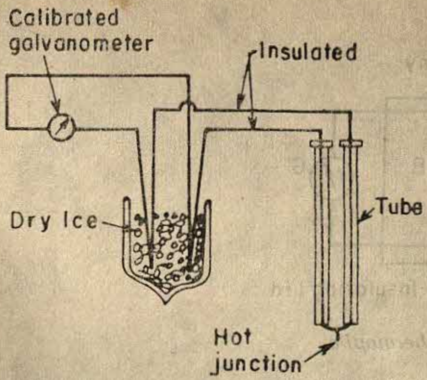


Fig. 4.16 *Thermoelectric thermometer*

about 1600°C. As a temperature measuring device, a thermoelectric thermometer has many merits.

(i) The thermometer measures the temperature of the hot junction. Since the size of this junction can be made extremely small, it is particularly suitable for measuring temperatures in very small regions or cavities.

(ii) It is very accurate. The copper-constantan thermocouple, for instance, produces an emf of 40 microvolts per degree temperature difference between its junctions. This emf can be easily measured with an accuracy of 1 microvolt. The temperature difference can, therefore, be measured with an accuracy of about 1/40 degree.

(iii) Because of its small mass and low thermal capacity, the test junction rapidly attains thermal equilibrium with the surroundings. It can, therefore, be used to detect variations in temperature of animals, insects, etc.

(b) *Thermopile* The thermopile is a device used for measuring very small temperature differences of the order of 0.001°C or for detecting and measuring the intensity of heat radiations. The emf generated by a thermocouple is very small, of the order of a few millivolts. The sensitivity of the device can be considerably increased if a number of thermocouples are connected in series so that the emfs are added up. Such a combination is called a thermopile.

Figure 4.17 shows a bismuth-antimony thermopile. A single thermocouple of bismuth and antimony can produce a comparatively large emf for a small difference in temperatures of the junctions. A series combination of such couples can be made so sensitive that it can detect heat radiations from a candle several metres away. One set of

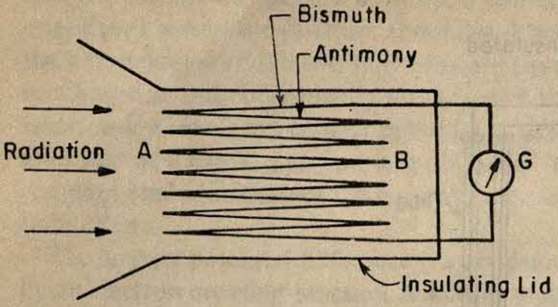


Fig. 4.17 *A bismuth-antimony thermopile*

junctions (A) is exposed to heat radiations and the other set (B) is protected by an insulating lid as shown in the figure. The exposed ends are blackened to increase the absorption of radiation falling on them. The deflection of the galvanometer is proportional to the intensity of the incident radiation.

(c) *Thermoelectric Refrigerator and Generator* We have seen that thermoelectric effects are reversible. So they can be used as heat engines and heat pumps (refrigerators). In a heat engine, thermoelectric effects can be used to absorb net heat and do useful work. In a refrigerator, they can be used to remove heat by doing work on the working substance (see Chapter 11 of the Class XI textbook). Thus Peltier effect can be used in a thermoelectric refrigerator. If a current is passed through a circuit with two junctions which are at different temperatures, one finds that heat is removed from one junction and liberated at the other. If the net result is that some heat is removed from the system, one has a refrigerator. Such a refrigerator obviously has no motors or moving parts and is, therefore, noise-free. Because of the low efficiency, such refrigerators are not in commercial use.

The Seebeck effect can be used in a thermoelectric generator. Extensive research is being carried out, especially in the Soviet Union, to develop generators which have a high thermo power, low resistivity (to minimize Joule losses) and low thermal conductivity (to enable large temperature differences to be maintained). The most suitable materials, from these points of view, are based on the semiconductor $BiTe_3$. In remote areas, thermoelectric generators are used in radio receivers and transmitters.

Example 4.13 The thermo emf generated in a thermocouple, whose cold junction is at 0°C and hot junction is at t°C is given by $\mathcal{E} = at$

$+ bt^2$, where a and b are constants of the couple. At two known temperatures 20°C and 240°C, the emf's generated by this couple are respectively 791 μV and 9482 μV. (i) Determine the constants a and b. (ii) At what value of t is $\mathcal{E} = 4160\ \mu V$? (iii) What is the neutral temperature of the couple?

Solution:

$$\mathcal{E} = at + bt^2$$

Given
$$\mathcal{E}_1 = 791\ \mu V, \quad t_1 = 20°C$$
$$\mathcal{E}_2 = 9482\ \mu V, \quad t_2 = 240°C$$

(i) Therefore, we have

$$791 = 20a + (20)^2 b$$

and
$$9482 = 240a + (240)^2 b$$

Solving these two equations simultaneously for a and b, we have

$$a = 42.8\ \mu V/°C$$

and

$$b = -0.012\ \mu V/(°C)^2$$

(ii) Using these values of a and b and $\mathcal{E} = 4160\ \mu V$ in the equation $\mathcal{E} = at + bt^2$, we have

$$4160 = 42.8t - 0.012t^2$$

This quadratic equation is solved for t. The positive root is

$$t = 100°C$$

(iii) The neutral temperature t_N is the value of t when $d\mathcal{E}/dt = 0$, i.e. $\mathcal{E}$ attains a stationary value. Differentiating the equation $\mathcal{E} = at + bt^2$ and setting $d\mathcal{E}/dt = 0$, we get

$$0 = a + 2bt$$

When this happens, $t = t_N$, hence

$$0 = \text{a} + 2bt_N$$

or
$$t_N = -\frac{a}{2b}$$

$$= -\frac{42.8}{2 \times -0.012} = 1783\ °C$$

EXAMPLE 4.14 A copper-constantan thermocouple produces an emf of 40 μV per degree Celsius in the linear range of temperatures. A galvanometer of resistance 10 Ω and capable of detecting currents of the order of 1 μA is employed. What is the smallest temperature that can be detected by this galvanometer?

Solution:

$$I = 1\mu\text{A} = 1\times10^{-6}\ \text{A}$$

$$R = 10\ \Omega$$

The smallest potential difference that can be detected by the galvanometer is

$$V = IR = 1\times10^{-6}\times10 = 10^{-5}\ \text{V}$$

If δt is the smallest difference in temperatures of the cold and hot junctions, the corresponding thermo emf is

$$40\ \mu\text{V}\times\delta t = 40\times10^{-6}\,\delta t$$

Thus

$$40\times10^{-6}\,\delta t = 10^{-5}$$

or

$$\delta t = \frac{10^{-5}}{40\times10^{-6}} = 0.25\ ^\circ\text{C}$$

SUPPLEMENT

RELATION BETWEEN SEEBECK, PELTIER AND THOMSON COEFFICIENTS

As a result of the absorption and evolution of heat energy at the junctions (Peltier effect) and all along the conductors (Thomson effect), a total Seebeck emf is produced which sends a current round the thermocouple.

Consider a thermocouple of two metals A (say copper) and B (say iron) having junctions 1 (cold) and 2 (hot) kept at absolute temperatures T and $(T + \mathrm{d}T)$. As a result let a current I (Seebeck effect) flow for, say, t seconds in the direction shown in Fig. 4.18. The charge transported by the current during t seconds is $q = It$. If π and $(\pi + \mathrm{d}\pi)$ are the Peltier coefficients at temperatures T and $(T + \mathrm{d}T)$ respectively and σ_A and σ_B are the Thomson coefficients of metals A and B respectively, then

$$\text{Peltier energy absorbed at junction 2} = q\pi_2 = q(\pi + \mathrm{d}\pi) \qquad \text{(i)}$$

$$\text{Peltier energy evolved at junction 1} = q\pi_1$$

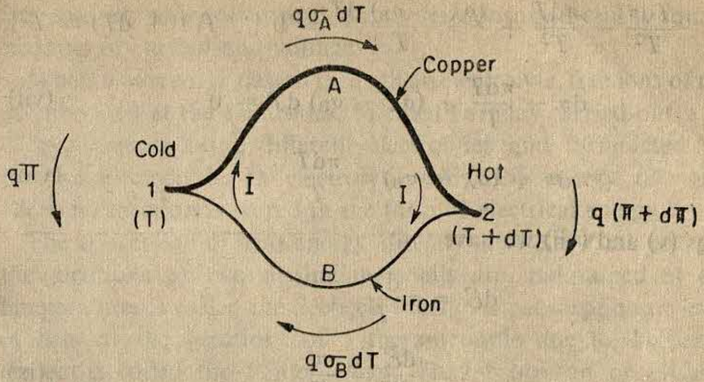


Fig. 4.18 *Relation between S, π and σ*

or Peltier energy absorbed at junction 1 $= -q\pi_1 = -q\pi$ (ii)

Thomson energy absorbed in metal A $= q\sigma_A \, dT$ (iii)

Thomson energy evolved in metal B $= q\sigma_B \, dT$

or Thomson energy absorbed in metal B $= -q\sigma_B \, dT$ (iv)

Thus, for the whole circuit, energy is absorbed at the hot junction (Peltier effect) and also along the two conductors (Thomson effect), while energy is evolved at the cold junction. If $d\mathcal{E}$ is the thermo emf of the couple, the total energy produced is $q \, d\mathcal{E}$ which is equal to the sum of the energies absorbed in the various parts of the circuit. Thus, adding (i) to (iv) and equating the sum to $q \, d\mathcal{E}$ we have

$$qd\mathcal{E} = q(\pi + d\pi) - q\pi + q\sigma_A dT - q\sigma_B dT$$

or
$$d\mathcal{E} = d\pi + (\sigma_A - \sigma_B)dT \quad \text{(v)}$$

or
$$d\mathcal{E} = d\pi + \sigma_{AB} \, dT \quad \text{(vi)}$$

where $\sigma_{AB} = \sigma_A - \sigma_B$.

Now we know that the Peltier and the Thomson effects are reversible. Hence a thermocouple may be compared to a reversible heat engine and the second law of thermodynamics may be applied. We have seen that the energy absorbed at $(T + dT)$ and given out at T due to Peltier effect are $q(\pi + d\pi)$ and $q\pi$ respectively and the energy absorbed due to Thomson effect is $q(\sigma_A - \sigma_B) \, dT$. Hence from the second law of thermodynamics we have

$$\frac{q(\pi + d\pi)}{(T + dT)} - \frac{q\pi}{T} + \frac{q(\sigma_A - \sigma_B) \, dT}{T} = 0$$

or
$$\frac{\pi T + Td\pi - \pi T - \pi dT}{T(T + dT)} + \frac{(\sigma_A - \sigma_B) \, dT}{T} = 0$$

or $$\frac{T\mathrm{d}\pi}{T^2} - \frac{\pi \mathrm{d}T}{T^2} + \frac{(\sigma_A - \sigma_B)\,\mathrm{d}T}{T} = 0 \quad [\because T(T+\mathrm{d}T) \simeq T^2]$$

or $$\mathrm{d}\pi - \frac{\pi \mathrm{d}T}{T} + (\sigma_A - \sigma_B)\,\mathrm{d}T = 0 \tag{vii}$$

or $$\mathrm{d}\pi + (\sigma_A - \sigma_B) = \frac{\pi \mathrm{d}T}{T} \tag{viii}$$

From Eqs (v) and (vii) we have

$$\mathrm{d}\mathcal{E} = \pi \frac{\mathrm{d}T}{T}$$

or $$S = \frac{\mathrm{d}\mathcal{E}}{\mathrm{d}T} = \frac{\pi}{T} \tag{ix}$$

which is relation (4.13). Equation (ix) gives

$$\pi = TS = T\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}T}$$

Thus the Peltier coefficient (π) at any temperature for any two metals forming a thermocouple is equal to the product of the absolute temperature (T) and the thermoelectric power $S(=\mathrm{d}\mathcal{E}/\mathrm{d}T)$ at that temperature. This gives us a practical method for measuring π.

Now from Eq. (vii) we have

$$\sigma_A - \sigma_B = -\frac{\mathrm{d}\pi}{\mathrm{d}T} + \pi\frac{\mathrm{d}T}{T}$$

$$= -T\frac{\mathrm{d}}{\mathrm{d}T}\left(\frac{\pi}{T}\right)$$

Using Eq. (ix) in this equation we get

$$\sigma_A - \sigma_B = -T\frac{\mathrm{d}}{\mathrm{d}T}\left(\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}T}\right)$$

or $$\sigma_{AB} = -T\frac{\mathrm{d}^2\mathcal{E}}{\mathrm{d}T^2} = -T\left(\frac{\mathrm{d}S}{\mathrm{d}T}\right) \tag{x}$$

which is relation (4.14). Thus the Peltier and Thomson coefficients are expressible in terms of the Seebeck coefficient.

SUMMARY

The heating of a conductor by the flow of an electric current through it is called Joule heating. This effect is thermodynamically irreversible. The amount of heat energy $= I^2Rt$. Electrical power is the rate at which electrical energy is consumed and is equal to the product of current and voltage.

When a current is passed through an electrolyte, free ions of elements are liberated at the electrodes. Michael Faraday carried out a number of experiments using different electrolytes and formulated the two laws of electrolysis. In electrolytic cells, the energy released in a chemical reaction is stored in the form of electrical energy.

The conversion of heat energy directly into electrical energy when the junctions of two dissimilar metals are maintained at different temperatures is called the Seebeck effect. The absorption or evolution of heat at the junctions of a thermocouple due to the passage of current is called the Peltier effect. The absorption or evolution of heat due to the flow of current in an unequally heated single conductor is called the Thomson effect.

EXERCISES

A. Short-Answer Questions

1. State Joule's law of the heating of a conductor.
2. Define electric power and state its SI unit.
3. Distinguish between kilowatt and kilowatt hour.
4. State Faraday's laws of electrolysis.
5. What do you understand by the electrochemical equivalent of a metal?
6. What is Faraday's constant? State its value.
7. State the Seebeck effect.
8. State the Peltier and Thomson effects.
9. What is the order of magnitude of thermo emf?
10. What do you understand by the thermoelectric power of a thermocouple?

B. Long-Answer Questions

1. From first principles, obtain the expression for the heat developed in a resistor by the passage of an electric current through it.
2. The rate P of Joule heat is given by

   $$P = I \times V$$

   Where I is the current in the conductor and V, the potential difference across its ends. Does this relation hold for a conductor that does not obey Ohm's law?
3. Using Ohm's law ($V = IR$), the rate of Joule heating can be written as either $P = I^2R$ or $P = V^2/R$. The equation $P = I^2R$ seems to suggest that the rate of Joule heating is reduced if R is decreased, but the equation $P = V^2/R$ seems to suggest just the opposite. How do you reconcile this apparent paradox?
4. Is the filament resistance lower or higher in a 200 W light bulb than in a 100 W light bulb? Both bulbs are designed to operate on 220 V.

5. How does the conduction of electricity in a metallic conductor differ from that in an electrolyte?
6. State Faraday's laws of electrolysis. How will you experimentally verify them?
7. Define electrochemical equivalent of an element. What is the meaning of the statement 'the electrochemical equivalent of copper is 0.000329 g C^{-1}'? How is it experimentally determined?
8. Describe a copper voltameter and explain how it can be used to measure current. What is the accuracy of its measurement?
9. Describe two important applications of electrolysis.
10. Describe the construction and the chemical action of:
    (i) Daniell cell and (ii) Leclanché dry cell.
11. Distinguish between primary and secondary cells. Describe the chemical action of a lead-acid accumulator when: (a) it is used to send a current through an external circuit, and (b) it is being charged.
12. What are fuel cells? What are their advantages over electrolytic cells?
13. Describe the working of a hydrogen-oxygen fuel cell.
14. What is the Seebeck effect? How can it be experimentally demonstrated?
15. What is a thermocouple? Graphically sketch the variation in the thermo emf generated by a couple as the temperature of the hot junction is gradually increased, the temperature of the cold junction being maintained at 0°C. What is the neutral temperature and the temperature of inversion of a couple?
16. Describe clearly how a potentiometer is calibrated to measure an emf of the order of 1 m V.
17. What is a thermopile? How can it be used for measuring the temperature of a hot body? List two advantages of using a thermocouple as a temperature measuring device.
18. What is Peltier effect? How does it differ from Joule's heating effect?
19. State and explain Thomson's effect.
20. Define the Seebeck, Peltier and Thomson coefficients in thermoelectricity and deduce the relations between them.

C. Multiple-Choice Questions

Choose the correct answer or answers from the given alternatives.

1. Choose the WRONG statement. The amount of heat developed in a conductor is director proportional to

   (a) the resistance of the conductor
   (b) the current flowing in the conductor
   (c) the square of the current in the conductor
   (d) the time for which the current is passed.

2. An electric heater has a rating of 220 V, 1100 W. How much electrical energy in kilowatt hours will it consume if used continuously for four hours?

   (a) 2.2 (b) 4.4
   (c) 6.6 (d) 8.8

3. What is the resistance of the element of the heater in Question 2?

(a) 22 Ω　　(b) 44 Ω
(c) 66 Ω　　(d) 88 Ω

4. An electric kettle has two coils. When one coil is connected to the ac mains, water in the kettle boils in 10 min. When the other coil is used, the same quantity of water takes 45 min to boil. How long will it take for the same quantity of water to boil if the two coils are connected in parallel?

(a) 6 min　　(b) 12 min
(c) 18 min　　(d) 24 min

5. Referring to Question 4, how long will it take for the same quantity of water to boil if the two coils are connected in series?

(a) 5 min 37.5 s　　(b) 6 min 37.5 s
(c) 7 min 37.5 s　　(d) 8 min 37.5 s

6. Three bulbs of ratings 40, 60 and 100 W are designed to work on a 220 V mains. Which bulb will burn most brightly if they are connected in series across 220 V mains?

(a) 40 W bulb　　(b) 60 W bulb
(c) 100 W bulb　　(d) All will burn equally brightly

7. Three equal resistors, connected in series with a battery, together dissipate 10 W of power. What will be the power dissipated if the same resistances are connected in parallel across the same battery?

(a) 10/3 W　　(b) 10 W
(c) 30 W　　(d) 90 W

8. Choose the correct statement. The masses of different substances liberated in electrolysis by the same quantity of electricity are proportional to their relative:

(a) atomic masses
(b) valencies
(c) ratios of atomic mass and valency
(d) products of atomic mass and valency

9. How many coulombs of electricity must pass through acidulated water to liberate 22.4 litres of hydrogen at STP?

(a) 1.6×10^{-19}C　　(b) 1 C
(c) 96500 C　　(d) 6.24×10^{28} C

10. During electrolysis, the current in the electrolyte is maintained by the flow of

(a) electrons only　　(b) negative ions only
(c) positive ions only　　(d) both positive and negative ions

11. For what time must a constant current of 2A pass through a copper voltameter to deposit 1 g of copper? The ece of copper is 0.00033 gC^{-1}.

(a) about 20 min　　(b) about 25 min
(c) about 30 min　　(d) about 35 min

12. An ammeter is suspected of giving inaccurate readings. In order to check the ammeter, it is connected in series with a silver voltameter and a steady current is passed for one hour. The ammeter indicates 0.54 A and 2.0124 g of silver are deposited. What is the error in the ammeter reading? The ece of silver is 0.001118 gC^{-1}.

(a) −0.01 A	(b) +0.02 A
(c) −0.03 A	(d) +0.04 A

13. A battery of emf 3 V and internal resistance 1.0 Ω is connected to a copper voltameter and a current of 1.5 A flows in the circuit. What is the resistance of the voltameter?

(a) Zero	(b) 1.0 Ω
(c) 1.5 Ω	(d) 2.0 Ω

14. A copper voltameter is connected in series with a heating coil of resistance 10 Ω. A steady current flows in the circuit for 20 min and deposits 0.99 g of copper. If the ece of copper is 0.00033 gC^{-1}, the amount of heat generated in the coil in this time is

(a) 25000 J	(b) 50000 J
(c) 75000 J	(d) 100000 J

15. Choose the correct statement. The thermo emf of a thermocouple
    (a) is independent of the nature of the pair of metals used in it
    (b) depends upon the temperature of the cold junction alone
    (c) depends upon the temperature of the hot junction alone
    (d) depends upon the temperature difference between the two junctions as well as the temperature of the cold junction.

16. A copper-constantan thermocouple produces an emf of 40 μV per degree Celsius in the linear range of temperatures. A galvanometer of resistance 10 Ω and capable of detecting currents of the order of 1μA is employed. What is the smallest temperature that can be detected by this galvanometer?

(a) 1/8°C	(b) 1/4°C
(c) 1/2°C	(d) 1°C

17. The thermo emf E (in volts) of a certain thermocouple is found to vary with θ (in °C) in accordance with the relation

$$E = 40\,\theta - \frac{\theta^2}{20}$$

where θ is the temperature of the hot junction; the cold junction being kept at 0°C. What is the neutral temperature of the couple?

(a) 100°C	(b) 200°C
(c) 300°C	(d) 400°C

D. Numerical Exercises

1. Heat is developed in a resistor at a rate of 90 W when the current is 3 A. What is the resistance?
2. What voltage drop will there be across a 2 kW electric heater element whose resistance when hot is 20 Ω?

3. A 1000 W immersion heater is placed in a pot containing 2 litre of water at 20°C. How long will it take to heat the water to boiling temperature if 20% of the available energy is lost to the surroundings?
4. A 1000 W immersion heater is placed in a copper geyser of mass 2 kg containing 20 litre of water at 20°C. How long will it take to heat the water to 70°C? Specific heat of water is 4200 J kg^{-1} K^{-1} and of copper is 390 J kg^{-1} K^{-1}.
5. A electric fan has a rating of 220 V, 40 W. Calculate (i) the resistance of the fan, and (ii) the cost of using it for 6 hours each day for 30 days at 30 paise per unit of electrical energy.
6. An electric kettle has two coils. When one coil is connected to the ac mains, water in the kettle boils in 10 minutes and when the other coil is used the same quantity of water takes 45 minutes to boil. How long will it take for the same quantity of water to boil if the two coils are connected in: (i) series and (ii) parallel?
7. At room temperature, the thermo emf generated in a thermocouple is 50 μV per degree Celsius. Assuming a linear relationship between the emf and temperature, compute the smallest temperature difference that can be detected with a thermocouple using a galvanometer of 20 Ω resistance and capable of detecting current as low as 1 μA.
8. The emf E (measured in microvolts) of a certain thermocouple is found to vary with t in accordance with the relation

$$E = 40t - \frac{t^2}{20}$$

where t is the temperature of the hot junction, the cold junction being kept at 0°C. What is the neutral temperature of the couple?
9. For what time must a constant current of 2 A pass through a silver voltameter to deposit 1.12 g silver? The electrochemical equivalent of silver is 0.00112 g C^{-1}.
10. A thin metal plate, 5 cm×4 cm, is to have a layer of copper 0.1 mm thick deposited on both sides. If a current of 1 A is used, how long will this take? The electrochemical equivalent of copper is 0.0003 g C^{-1} and its density is 9000 kg m^{-3}.
11. A battery of emf 3 V and internal resistance 1.0 Ω is connected to a copper voltameter and a current of 1.5 A flows in the circuit. Calculate the resistance of the voltameter.
12. A current is passed through a silver nitrate solution for an hour and the mass of silver liberated is determined. How long should a current of the same strength be passed through a solution of ferric chloride to liberate iron of the same mass?
13. A copper voltameter is connected in series with a heating coil of resistance 10 Ω. A steady current flows in the circuit for 20 minutes and deposits 0.99 g of copper. Calculate the amount of heat generated in the coil in this time. The electrochemical equivalent of copper is 0.00033 g C^{-1}.
14. The walls of a closed cubical box of edge 50.0 cm are made of a material of thickness 1.00 mm and thermal condutivity 1.64×10^{-1} J s^{-1} $m^{-1}K^{-1}$. The interior of the box is maintained at 100°C above the outside temperature by a heater inside the box and connected across a 400 V dc source. Calculate the resistance of the heater.

15. What must be the wattage of a bulb which will glow with maximum brightness when it is connected across a 4.5 V battery of internal resistance 2 Ω? (Hint: For maximum dissipation of energy, the resistance of the bulb must be equal to the internal resistance of the battery.)
16. A hall is lighted by 12 bulbs, each rated 100 W, 200 V by a dynamo of emf 215 V. Find the resistance of the leads. What is the percentage of energy lost in the leads?
17. The maximum power rating of a resistance box is 1.00×10^{-4} watt per ohm. Find the limiting safe voltage that can be applied to the box when (a) a 1.5 Ω and (b) a 2500 Ω resistor is being used.
18. The resistance of the nichrome filament of a heater measured at room temperature of 15°C is 32 Ω. When connected to 230 V mains, the filament dissipates 1250 W. If the temperature coefficient of resistivity of nichrome is $4.00 \times 10^{-5}/°\mathrm{C}$, find the working temperature of the filament.
19. The thermelectric power $(d\mathcal{E}/dT)$ for iron against lead at t°C is given by $1734 - 4.87\,t$ and that for copper by $1376 + 0.95\,t$. Find the thermo emf of a copper-iron thermocouple with junctions at 20°C and 200°C.
20. If 1.00 g of hydrogen on burning to form water liberates 1.47×10^{3} J of energy, find the back emf of a water voltameter when it is connected to a 2.0 V accumulator. The ece of hydrogen is 1.04×10^{-5} g C^{-1}.

5

MAGNETIC EFFECTS OF CURRENT

5.1 INTRODUCTION

The science of electricity evolved from the observation of the Greek philosopher, Thales, in the sixth century BC, that a rubbed piece of amber will attract pieces of straw. The science of magnetism goes back to the observation that naturally occurring 'stones' (e.g. magnetite) will attract iron. The two sciences developed independently until 1820, when Hans Christian Oersted discovered the connection between them. He showed that a current in a wire can also produce magnetic effects, namely, that it can change the orientation of a compass needle. This gave birth to a new science, electromagnetism, which was further developed by many workers; the most notable contributions coming from Michael Faraday (1791–1867) and James Clerk Maxwell (1831–1879). Oersted's experiment clearly established that a current carrying wire exerts a force on a magnet such as a compass needle. Faraday argued that, from Newton's third law, the reverse should also be true, i.e. the magnet should also exert an equal and opposite force on the wire carrying current. It was Maxwell who put the laws of electromagnetism in the form of a set of simple equations, called Maxwell's equations that describe the behaviour of the electric and magnetic fields.

5.2 OERSTED'S EXPERIMENT

Hans Christian Oersted, Professor of physics at Copenhagen in Denmark, was the first to discover the connection between electric current and magnetism. Let us perform the famous experiment he performed in the year 1820. We need only a compass needle, a length of a wire, a battery and a switch (or key). Set up the circuit as shown in Fig. 5.1. A section AB of the wire is set parallel to the magnetic needle [Fig. 5.1(a)]. When the key is not inserted, the wire and the needle are both pointing North-South. The moment the key is inserted the magnetic needle swings to a new position as shown in Fig. 5.1(b).

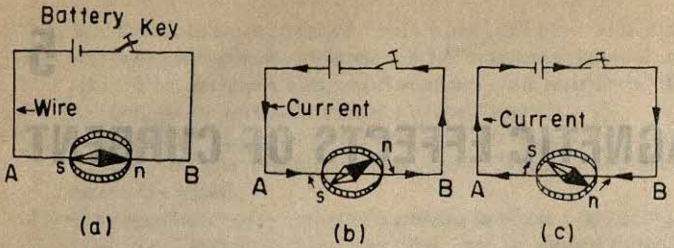


Fig. 5.1 *Oersted's experiments*

When the current in the circuit is reversed the needle deflects in the opposite direction [Fig. 5.1(c)].

It is important to realise that the deflection of the needle is due to the magnetic field and not due to any electric field. We know that there exists an electric field around a charge. When a current flows through a wire, there is no net charge on the wire. The electrons move from one end of the wire to the other, and the number of electrons entering one end is equal to the number of electrons leaving the other end, so that the wire stays electrically neutral as the charges move through it. Hence there is no net electric field around the wire. The needle is deflected by the magnetic field. Thus *a moving charge*, i.e. current, *produces a magnetic field.*

We may go a step further and make a general statement that a moving charge is a source of magnetic field. When there is no current in the wire, the electrons move about in the wire due to their thermal motion. Each electron acts as the source of a magnetic field. Since the thermal motions are random (or undirected) due to collisions, these fields cancel out. There is no net magnetic field as there is no net flow of charge in any direction.

5.3 BIOT-SAVART LAW AND MAGNETIC FIELD DUE TO A CURRENT

The next problem is to obtain the quantitative relationship between the current in any conductor and the resulting magnetic field at a point in the space around it. Each moving charge in a conductor contributes to the magnetic field. The net magnetic field at a point is the sum of all the contributions of individual charges. Since the number of charges is enormously large, this procedure is obviously hopeless. In physics, whenever we have to deal with an infinitely large number of contributions, we often use a simplified procedure which we shall illustrate by the following simple example.

Let us consider the problem of calculating the magnetic field **B** at a point P due to a current I flowing in a wire bent into any arbitrary shape (Fig. 5.2). Rather than deal with an infinite number of individual charges, we divide the wire into small elements (called current elements) placed end to end. Let the length of each infinitesimally small element be d**l**. In 1820 Biot and Savart succeeded in finding the contribution of the magnetic field at a point due to such a current element. This was stated in the form of a law called the Biot-Savart law. According to this law, the magnitude of the magnetic field d**B** at a point P due to a current element d**l** is given by

$$\mathrm{d}B = \frac{\mu_0}{4\pi}\,\frac{I\,\mathrm{d}l \sin\theta}{r^2} \tag{5.1}$$

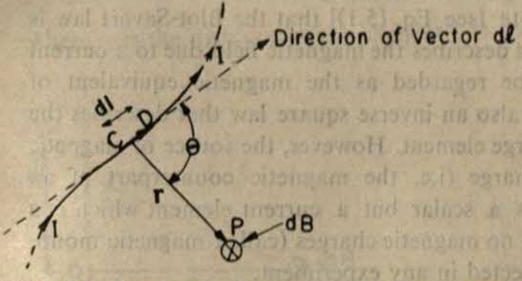


Fig. 5.2 *Biot-Savart Law. Magnetic field contribution* *d*B *at a point P due to a current element d*l

where **r** is a displacement vector from the element say CD to P, the observation point, and θ is the angle between this vector and element length d**l**. The length of the element is regarded here as a vector d**l** whose magnitude is d**l** and whose direction is that of the tangent to the conductor at that point (shown dashed in Fig. 5.2). I is the current in the conductor and μ_0 is a constant called the permeability of free space (vacuum). Equation (5.1) gives only the magnitude of d**B**. The direction of d**B** is that of the vector (d**l** × **r**). It is perpendicular to both d**l** and r. In Fig. 5.2, for example, d**B** at point P for the current element shown is directed into the page (shown by a cross, indicating the tail of the arrow) at right angles to the plane of the figure.

The law of Biot and Savart may be written in a vector form as

$$\mathrm{d}\mathbf{B} = \frac{\mu_0 I}{4\pi}\,\frac{(\mathrm{d}\mathbf{l}\times\mathbf{r})}{r^3} \tag{5.2}$$

This formulation reduces to that of Eq. (5.1) when expressed in terms of magnitudes. The law as formulated in Eq. (5.2) gives complete information about the magnitude as well as the direction of d**B**.

The resultant magnetic field **B** due to all the current elements of a conductor is then found by integrating (i.e. summing up the individual contributions), i.e.

$$\mathbf{B} = \int d\mathbf{B}$$

It may be mentioned that it is possible to give a rigorous theoretical proof of this law which is beyond the scope of this book. Biot and Savart, however, arrived at the law only empirically basing their arguments on the experimental results of the magnetic effect of current. We shall illustrate this law by using it to calculate the magnetic field due to some simple current distributions. In each case, the results are found to be in complete agreement with experimental observations.

It is interesting to note [see Eq. (5.1)] that the Biot-Savart law is an inverse square law that describes the magnetic field due to a current element. This law may be regarded as the magnetic equivalent of Coulomb's law, which is also an inverse square law that describes the electric field due to a charge element. However, the source of magnetic field is not a magnetic charge (i.e. the magnetic counterpart of an electric charge) which is a scalar but a current element which is a vector. In fact, till today, no magnetic charges (called magnetic monopoles) have ever been detected in any experiment.

In fact, all known steady magnetic fields are produced either by electric currents or by atomic magnetic moments associated with angular momenta of charged particles (see Chapter 6). Magnetic fields are also produced by time-dependent electric fields, e.g. due to accelerated electric charges, as we will see in Chapter 9.

The SI Unit of Magnetic Field

The SI unit of magnetic field is called *tesla* (symbol T) after the name of the great Yugoslav scientist Nikola Tesla (1856–1943). It can be defined by Eq. (5.1) where the constant $\mu_0/4\pi$ is given the exact value 10^{-7} TmA^{-1}. Thus a tesla may be defined as follows: A small current element d**l** metre long, carrying a current of 1 A, produces, at a distance of 1 m from it and perpendicular to it, a magnetic field of 10^{-7} tesla.. This defines tesla in terms of current. In the next section, we will define tesla in terms of the force experienced by a moving charge (i.e. current). In chapter 9 we will find the relationship between the constant $(\mu_0/4\pi)$ and the constant $(4\pi\epsilon_0)$ (a constant used in electrostatics in chapter 1). It will be shown that the product of these two constants $(\mu_0\epsilon_0) = 1/c^2$ where c is the velocity of light in vacuum $= 3\times 10^8$ ms^{-1}. Thus this product has a fixed size and dimension.

Therefore, if one of them is chosen according to convenience, the other one is automatically fixed. We will discuss this in the next section.

We will now apply the Biot-Savart formula to obtain the magnetic field at a point due to some simple current distributions.

(a) *Magnetic Field due to an Infinitely Long Current Carrying Straight Wire* We will use the law of Biot and Savart to find **B** due to a current I in an infinitely long straight wire. Figure 5.3(a) shows a side view of the wire. We wish to find **B** at a point P distant r from the wire. The magnitude of the contribution d**B** due to a current element CD of length d**l** is given by

$$dB = \left(\frac{\mu_0}{4\pi}\right)\frac{I\,dl\sin\theta}{x^2}$$

where x is the distance of the element CD from point P.

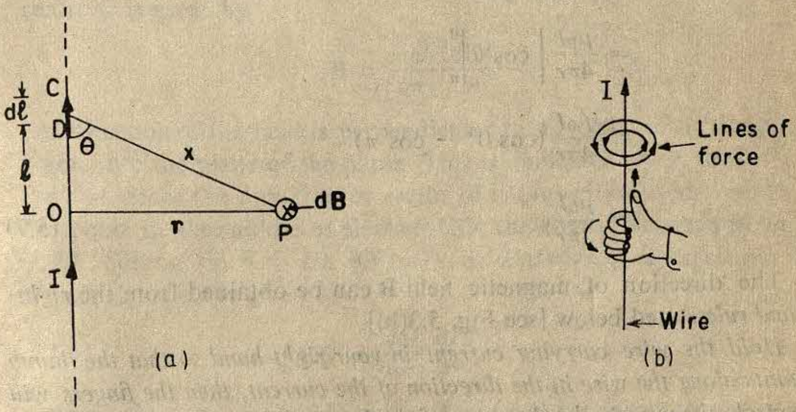


Fig. 5.3 *(a) Magnetic field due to an infinitely long straight wire, (b) magnetic field lines*

The directions of the contributions d**B** at point P for all elements are the same, namely, pointing into the page of the figure at right angles to the plane. The magnitude of the total field at point P is given by

$$B = \frac{\mu_0 I}{4\pi}\int_{l=-\infty}^{l=+\infty}\frac{\sin\theta\,dl}{x^2}$$

where l is the distance of the element from O, the foot of the perpendicular drawn from P on the wire.

It is clear from the figure that l, θ and r are related by

$$\frac{r}{x} = \sin\theta \qquad \text{or} \qquad x = \frac{r}{\sin\theta} \tag{5.3}$$

$$\frac{r}{l} = \tan\theta \qquad (5.4)$$

or

$$l = r\cot\theta \qquad (5.5)$$

Differentiating this equation we get

$$dl = -r\,\mathrm{cosec}^2\,\theta\,d\theta \qquad (5.6)$$

From Eq. (5.4) it is clear that when $l \to +\infty, \theta \to 0$ and when $l \to -\infty, \theta \to \pi$. Using Eqs (5.3) and (5.6) in the equation for B, we get

$$B = \frac{\mu_0 I}{4\pi}\int_{\pi}^{0} \sin\theta\,(-r\,\mathrm{cosec}^2\,\theta\,d\theta)\,\frac{\sin^2\theta}{r^2}$$

$$= -\frac{\mu_0 I}{4\pi r}\int_{\pi}^{0} \sin\theta\,d\theta$$

$$= \frac{\mu_0 I}{4\pi r}\left|\cos\theta\right|_{\pi}^{0}$$

$$= \frac{\mu_0 I}{4\pi r}(\cos 0^\circ - \cos\pi)$$

$$= \frac{\mu_0 I}{2\pi r} \qquad (5.7)$$

The direction of magnetic field **B** can be obtained from the *right-hand rule* stated below [see Fig. 5.3(b)].

Hold the wire carrying current in your right hand so that the thumb points along the wire in the direction of the current, then the fingers will encircle the wire in the direction of the lines of magnetic force. A line of force is a curve, the tangent to which at any point gives the direction of the magnetic field at that point. The lines of force due to the field of a long current-carrying wire are concentric circles in a plane perpendicular to the plane of the wire. The magnetic field at all points on a circle has the same value and its direction is perpendicular to the radius vector.

Note: Strictly speaking, the expression (5.7) holds only for an infinitely long wire which is an idealization. In practice, if $r \ll l$, the correction to Eq. (5.7) is of order of $(r/l)^2$ which may be ignored.

(b) *Magnetic Field due to a Current Carrying Circular Loop* We will now apply the Biot-Savart law to find the magnetic field due to a current carrying circular loop of wire. We consider the following two cases:

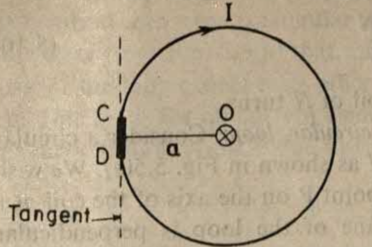


Fig. 5.4 *Magnetic field at the centre of a circular loop carrying current*

Case I—*Field at the centre of a circular loop*: Consider a loop (or a circular coil) of uniform wire carrying a current I as shown in Fig. 5.4. Let a be the radius of the loop. Let us assume that the loop lies in the plane of the paper. We wish to calculate the magnetic field at the centre O of the loop.

We divide the wire into a large number of current elements each of infinitesimally small length $\mathbf{dl}$. Consider one such element, say CD, of the loop. From Biot-Savart law, the magnetic field $\mathrm{d}\mathbf{B}$ (due to this current element) at the centre O is given by

$$\mathrm{d}\mathbf{B} = \frac{\mu_0 I}{4\pi}\frac{(\mathbf{dl}\times\mathbf{a})}{a^3}$$

The direction of the field is perpendicular to the plane of the loop and points into the plane of the paper. This is indicated by a cross ⊗ in Fig. 5.4. Since the direction of vector $\mathbf{dl}$ is tangential to the circle at the point in the middle of element CD, the angle θ between $\mathbf{dl}$ and $\mathbf{a}$ is 90°. Setting $\sin\theta = \sin 90^\circ = 1$ in the preceding equation the magnitude of $\mathrm{d}\mathbf{B}$ is given by

$$\mathrm{d}B = \frac{\mu_0 I}{4\pi}\frac{\mathrm{d}l}{a^2} \tag{5.8}$$

Now since the angle θ between $\mathbf{dl}$ and $\mathbf{a}$ is the same (equal to 90°) for all elements, the magnitude of the field due to each element is given by Eq. (5.8). Further, the direction of $\mathrm{d}\mathbf{B}$ due to each element is the same, pointing into the plane of the paper. Therefore, the resultant field at O is given in magnitude by

$$B = \int \mathrm{d}B = \frac{\mu_0 I}{4\pi a^2}\int \mathrm{d}l$$

But $\int \mathrm{d}l$ is just the total length of the loop of wire which is equal to $2\pi a$. Hence.

$$B = \frac{\mu_0 I}{4\pi a^2}(2\pi a) = \frac{\mu_0 I}{2a} \tag{5.9}$$

If instead of a single loop of wire, there is a coil of N turns wound one over another, then clearly the total length of the wire is $2\pi aN$

and the magnetic field at the centre of the coil is given by

$$B = \frac{\mu_0 I N}{2a} \tag{5.10}$$

where a is the mean radius of the coil of N turns.

Case II—*Field on the axis of a circular loop*: Consider a circular loop of radius a carrying a current I as shown in Fig. 5.5(a). We wish to calculate the magnetic field at a point P on the axis of the coil at a distance r from its centre O. The plane of the loop is perpendicular to the plane of the paper. As before, we divide the coil into a large number of elements, each of length d**l**. The vector d**l** of the current element at the top C of the loop points out of the page at right angles to it. Therefore, the angle θ (not shown in the figure) between d**l** and **R** (the distance vector from the centre of the element at C to point P) is 90°. Hence the magnitude of field d**B** at point P due to the element at C is given by (from Biot-Savart law)

$$dB = \frac{\mu_0}{4\pi} \frac{I\,dl}{R^2}$$

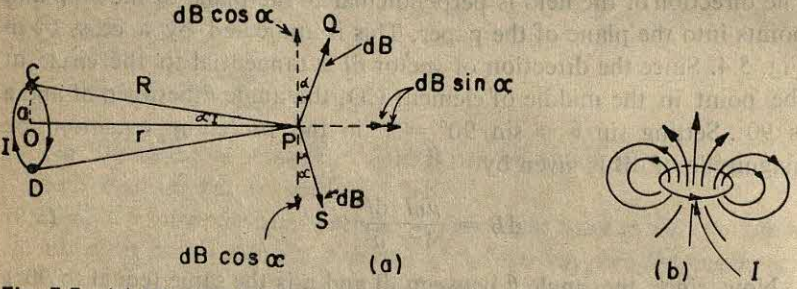


Fig. 5.5 *(a) Magnetic field on the axis of a circular coil carrying current, and (b) magnetic field lines*

To find the direction of d**B** let us recall that vector d**l** is perpendicular to the plane of the paper. Since the direction of field d**B** is perpendicular to both d**l** and **R**, and lies in the plane of the paper, it is clear that d**B** lies in the plane of paper and is perpendicular to **R**. This direction is shown as PQ in the figure. Let α be the angle between CP and the axis OP of the loop. Let us resolve vector d**B** into two rectangular components, one along the axis of the coil ($dB \sin \alpha$) and the other at right angles to this axis ($dB \cos \alpha$).

From symmetry, there is a point D opposite to point C where the current element points inside the paper at right angles to it. This follows from the direction (anti-clockwise) of the current as shown in Fig. 5.5(a). It is clear from symmetry that the magnitude of d**B** at point P due to this element is the same as that due to the element at C, but the direction of d**B** is along PS as shown. This vector may also be resolved into two rectangular components $dB \sin \alpha$ and $dB \cos \alpha$. It is clear from Fig. 5.5(a) that the contributions perpendicular to the axis of the loop cancel each other while those parallel to the axis add up. This holds for all current elements of the loop. Hence, the total magnetic field at point P due to the loop is given by

$$B = \int dB \sin \alpha$$

Substituting for dB, we have

$$B = \int \frac{\mu_0}{4\pi} \frac{I\,dl}{R^2} \sin \alpha$$

Since μ_0 and I are constant, and R and α are constant for all points on the circular loop, we have

$$B = \frac{\mu_0}{4\pi} \frac{I \sin \alpha}{R^2} \int dl$$

Now, since $dl = 2\pi a$, where a is the radius of the loop, we have

$$B = \frac{\mu_0 I a \sin \alpha}{2R^2}$$

From Fig. 5.5(a), it is clear that

$$\sin \alpha = \frac{a}{R}$$

and

$$R = (r^2 + a^2)^{1/2}$$

Substituting for $\sin \alpha$ and R, we get

$$B = \frac{\mu_0 I a^2}{2(r^2 + a^2)^{3/2}} \qquad (5.11)$$

If there are N turns in the coil, then

$$B = \frac{\mu_0 N I a^2}{2(r^2 + a^2)^{3/2}} \qquad (5.12)$$

Equation (5.11) shows that the field falls off rapidly with distance r. Figure 5.5(b) shows the magnetic field lines due to a circular coil.

Notice that if $r = 0$ (at the centre of the loop), Eq. (5.11) reduces to Eq. (5.9) obtained for case I.

EXAMPLE 5.1 A horizontal overhead power line carries a current of 90 A in the east to west direction. Find the magnitude and direction of magnetic field due to the current 1.5 m below the line.

Solution: The magnitude of magnetic field is given by

$$B = \frac{\mu_0 I}{2\pi r}$$

where $I = 90$ A, $r = 1.5$ m and $\mu_0 = 4\times10^{-7}$ Tm A^{-1}. Substituting these values, we get

$$B = 1.2\times10^{-5}\ \text{T}$$

Applying the right-hand side rule, we find that the direction of **B** is towards the south.

EXAMPLE 5.2 Two concentric circular coils X and Y of radii 16 cm and 10 cm respectively lie in the same vertical plane containing the north-south direction. Coil X has 20 turns and carries a current of 16 A; coil Y has 25 turns and carries a current of 18 A. The sense of the current in X is anticlockwise, and in Y clockwise, for an observer looking at the coils facing the west. Give the magnitude and direction of the net magnetic field due to the coils at their common centre.

Solution: The magnitude of magnetic field at the centre of coil Y is given by

$$B_Y = \frac{\mu_0}{2}\,\frac{I_Y N_Y}{a_Y}$$

$$= \frac{4\pi\times10^{-7}}{2}\times\frac{18\times25}{(0.1)} = 9\pi\times10^{-4}\ \text{T}$$

Since the current in coil Y is clockwise, the direction of field B_Y is towards the west (see Fig. 5.6).

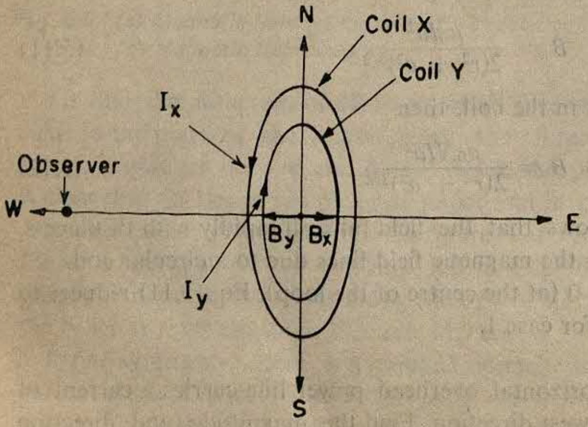


Fig. 5.6

The magnitude of magnetic field at the centre of coil X is given by

$$B_X = \frac{\mu_0}{2}\frac{I_X N_X}{a_X}$$

$$= \frac{4\pi\times 10^{-7}}{2}\times\frac{16\times 20}{0.16} = 4\pi\times 10^{-4}\ \text{T}$$

Since the current in coil X is anticlockwise, the direction of field B_X is towards the east as shown in Fig. 5.6. Since the two fields are collinear and oppositely directed, the magnitude of the resultant field = difference between the two fields and its direction is that of the bigger field. Hence the net magnetic field at the common centre is $5\pi\times 10^{-4}$ T = 1.6×10^{-3} T and is directed towards the west.

EXAMPLE 5.3 Helmholtz, an eminent scientist of the 19th century, showed that two narrow circular coils of the same radius (a) and carrying the same current (I) in the same direction can produce a uniform magnetic field over a small region around a point mid way between them. Show that the magnetic field over this small region is given approximately by

$$B = 0.72\,\frac{\mu_0 NI}{a}$$

Where N is the number of turns in each coil.

Solution: Consider a small region PQ = $2d$ near the point O which is midway between two coils X and Y, each of radius a. Thus OP = OQ = $a/2$ (see Fig. 5.7).

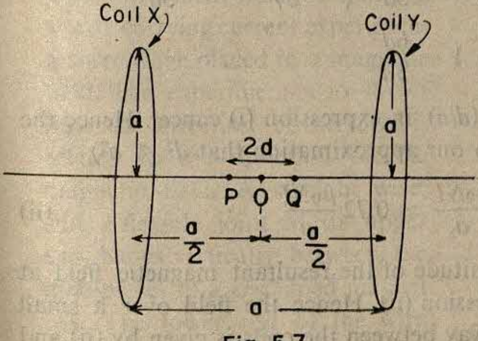


Fig. 5.7

Let us find the resultant magnetic field at point P due to the two coils. The magnetic field at point P due to coil X is [use Eq. (5.12) and put $r = (a/2 - d)$].

$$B_X = \frac{\mu_0 N I a^2}{2} \times \frac{1}{\left\{\left(\frac{a}{2} - d\right)^2 + a^2\right\}^{3/2}}$$

The magnetic field at point P due to coil Y is [setting $r = (a/2 + d)$ in Eq. (5.12)]

$$B_Y = \frac{\mu_0 N I a^2}{2} \times \frac{1}{\left\{\left(\frac{a}{2} + d\right)^2 + a^2\right\}^{3/2}}$$

Since the currents in the two coils are in the same direction, the fields B_X and B_Y act in the same direction so that the magnitude B of the resultant field is simply the sum of B_X and B_Y, i.e.

$$\begin{aligned} B &= B_X + B_Y \\ &= \frac{\mu_0 N I a^2}{2}\left[\frac{1}{\left\{\left(\frac{a}{2} - d\right)^2 + a^2\right\}^{3/2}} + \frac{1}{\left\{\left(\frac{a}{2} + d\right)^2 + a^2\right\}^{3/2}}\right] \\ &= \frac{\mu_0 N I a^2}{2} \cdot \frac{1}{\left(\frac{5a^2}{4}\right)^{3/2}}\left[\left(1 - \frac{4d}{5a}\right)^{-3/2} + \left(1 + \frac{4d}{5a}\right)^{-3/2}\right] \\ &= \frac{\mu_0 N I}{2a}\left(\frac{4}{5}\right)^{3/2} \times \left[1 + \frac{6d}{5a} + 1 - \frac{6d}{5a}\right] \qquad \text{(i)} \end{aligned}$$

where we have expanded the terms in the square bracket binomially and neglected terms of order $(d/a)^2$ and higher which is justified since $d \ll a$, i.e.

$$\left(1 - \frac{4d}{5a}\right)^{-3/2} \simeq 1 - \frac{4d}{5a} \times \left(-\frac{3}{2}\right) = 1 + \frac{6d}{5a}$$

and $$\left(1 + \frac{4d}{5a}\right)^{-3/2} \simeq 1 - \frac{6d}{5a}$$

Notice that terms of order (d/a) in expression (i) cancel. Hence the field at point P is given by (to our approximation that $d^2 \ll a^2$)

$$B = \left(\frac{4}{5}\right)^{3/2} \frac{\mu_0 N I}{a} \simeq 0.72 \frac{\mu_0 N I}{a} \qquad \text{(ii)}$$

It is evident that the magnitude of the resultant magnetic field at point Q is also given by expression (ii). Hence the field over a small region around a point midway between the coils is given by (ii) and is uniform over this region: such an arrangement to produce a nearly uniform magnetic field over a small region is known as Helmholtz coils.

EXAMPLE 5.4 A straight wire carrying a current of 12 A is bent into a semicircular arc of radius 2.0 cm as shown in Fig. 5.8(a). What is the direction and magnitude of **B** at the centre of the arc? Would your answer change if the wire were bent into a semicircular arc of the same radius but in the opposite manner, as shown in Fig. 5.8(b)?

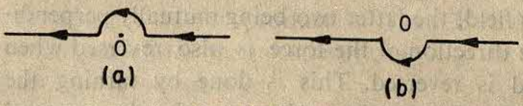


Fig. 5.8

Solution: The field at the centre O of the arc due to each straight portion of the wire is zero because point O lies on the line joining the straight portions. Thus the field at O is only due to half the circular loop and hence its magnitude is [see Eq. (5.9)]

$$B = \frac{1}{2} \cdot \frac{\mu_0 I}{2a}$$

$$= \frac{1}{2} \times \frac{4\pi \times 10^{-7} \times 12}{2 \times 0.02} = 1.9 \times 10^{-4}\ \mathrm{T}$$

Since the current in the arc is in the anticlockwise sense, the field is normal to the page and directed into it. If the wire is bent as shown in Fig. 5.8(b), the magnitude of **B** remains unchanged but its direction will now be normal to the page towards the reader, since the current round the loop is clockwise now.

5.4 FORCE ON A MOVING CHARGE IN A MAGNETIC FIELD

The experiment illustrated in Fig. 5.9 is designed to demonstrate that a wire carrying current experiences a force when placed in a magnetic field. The experiment also shows how the direction of the force is related to the direction of the magnetic field. A straight wire, with a flexible joint at its upper end, hangs vertically between the pole-pieces of a U-shaped magnet NS. The lower end of the wire dips into a small pool of mercury in a depression cut into a flat block of wood. When, the key is closed, a current flows in the

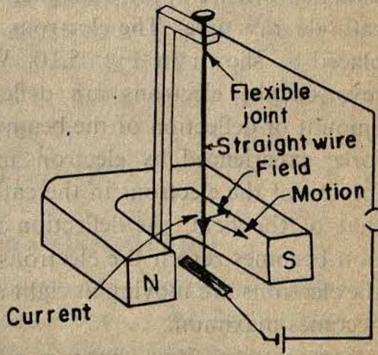


Fig. 5.9 *Force on a current carrying conductor in a magnetic field*

vertical wire downwards. The wire immediately experiences a force in the direction shown in the figure. When the current is reversed (by interchanging the terminals of the battery), the wire moves in the opposite direction. The direction of the force exerted on the wire is seen to be perpendicular to the direction of the current and the direction of the magnetic field, the latter two being mutually perpendicular to each other. The direction of the force is also reversed when the direction of the field is reversed. This is done by turning the magnet over so that its N-pole is at the position previously occupied by its S-pole.

It is misleading to say that the magnetic field exerts a force on the wire. In fact, a magnetic field exerts a force on any moving charge, be it in free space or in a metallic wire. Since the charges moving in a wire are confined to the body of the wire, the wire itself experiences a force. The fact that the field acts on charges moving freely can be established if we apply a magnetic field to the moving electrons in a cathode ray tube. The electrons are bent when a U-shaped magnet is placed as shown in Fig. 5.10. When the polarity of the magnet is reversed the electrons are deflected in the opposite direction. The amount of deflection of the beam is a measure of the magnitude of the force experienced by electrons moving in the field. If the direction of motion of the electrons in the cathode ray tube is changed relative to that of the field, the deflection of the beam also changes. The deflection becomes zero if the electrons move along a particular direction. If the electrons are moving at right angles to this direction, the deflection becomes maximum.

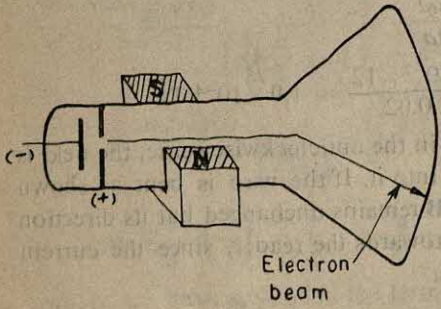


Fig. 5.10 *Deflection of electrons in a magnetic field*

Suppose a positive charge q is fired with a velocity $\mathbf{v}$ in a uniform magnetic field $\mathbf{B}$. Experiments tell us that the force $\mathbf{F}$ experienced by the charge depends not only on the magnitudes of $\mathbf{v}$ and $\mathbf{B}$ but also on the angle between them. Force **F** is found to be zero if **v** is parallel to **B**, i.e. if the angle θ between **v** and **B** is zero, but **F** is maximum if **v** and **B** are perpendicular to each other, i.e. if $\theta = 90^\circ$. Experiments also show that the direction of force **F** is always perpendicular to both **v** and **B**. These observations suggest that the force **F** must be expressed as a vector product of **v** and **B**. Since the magnitude of **F** depends also on the magnitude of charge q, we may express **F** by the following vector equation:

$$\mathbf{F} = q(\mathbf{v} \times \mathbf{B}) \tag{5.13}$$

The magnitude F of vector **F** is given by

$$F = qvB \sin \theta \tag{5.14}$$

It is clear from Eq. (5.13) that the direction of **F** is perpendicular to both **v** and **B** in accordance with experimental observation.

Figure 5.11 shows the relationship among the three vectors **B**, **F** and **v**. Vectors **B** and **v** lie in the x-y plane. The force **F** is at right angles to this plane and points along the z-axis. Notice that **F** is at right angles to **v** and also to **B**. Equation (5.13) and (5.14) are consistent with experimental observations which are:

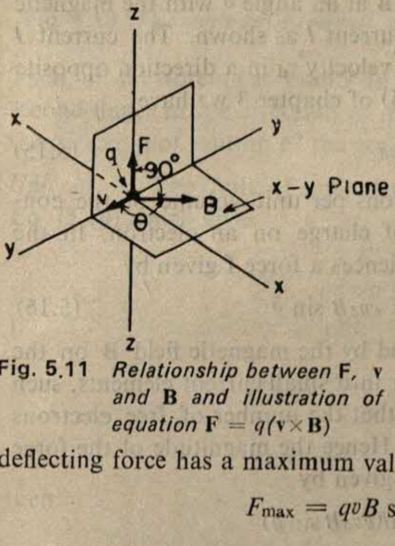


Fig. 5.11 *Relationship between* F, v *and* B *and illustration of equation* $\mathbf{F} = q(\mathbf{v} \times \mathbf{B})$

1. $\mathbf{F} = 0$, if $\mathbf{v} = 0$, i.e. a charge at rest does not experience any magnetic force.

2. $\mathbf{F} = 0$ if $\theta = 0$ or 180° i.e. the magnetic force vanishes if **v** is either parallel or antiparallel to the direction of **B**, and

3. $F = F_{\max}$ if $\theta = 90^\circ$, i.e. if **v** is at right angles to **B**, the deflecting force has a maximum value given by

$$F_{\max} = qvB \sin 90^\circ = qvB$$

Force on a Current-Carrying Conductor in a Magnetic Field

A current is an assembly of moving charges. In a current-carrying conductor, the free (conduction) electrons move in the body of the conductor. Because a magnetic field exerts a sideways force on a moving charge, we expect that it will exert a sideways force on a conductor carrying a current also.

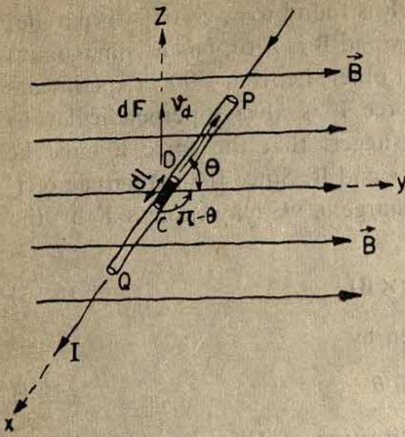


Fig. 5.12 *Force on a current-carrying conductor in a uniform magnetic field*

Figure 5.12 shows a conductor PQ of length l, cross-sectional area A placed in a uniform magnetic field **B** at an angle θ with the magnetic field lines. The conductor carries a current I as shown. The current I is due to the electrons drifting with velocity v_d in a direction opposite to that of the current. From Eq. (3.5) of chapter 3 we have

$$I = en_oAv_d \tag{5.15}$$

where n_o is the number of free electrons per unit volume of the conductor and e is the magnitude of charge on an electron. In the magnetic field **B** each electron experiences a force **f** given by

$$\mathbf{f} = e(\mathbf{v}_d \times \mathbf{B}) = ev_dB \sin\theta \tag{5.16}$$

In order to calculate the force exerted by the magnetic field **B** on the conductor, we divide the conductor into small current elements, such as CD, each of length dl. It is clear that the number of free electrons in an element of length dl is $n_oA\,\mathrm{d}l$. Hence the magnitude of the force experienced by a current element is given by

$$\mathrm{d}F = n_oA\,\mathrm{d}l(ev_dB\sin\theta)$$
$$= n_oeA\,\mathrm{d}lv_dB\sin\theta$$

Using Eq. (5.15) we have

$$\mathrm{d}F = I\mathrm{d}lB\sin\theta$$

In vector notation this reads

$$\mathrm{d}\mathbf{F} = I(\mathrm{d}\mathbf{l} \times \mathbf{B})$$

where d**l** is the vector length of the current element directed along the current. Since **B** is uniform over the length of the conductor, the total force **F** exerted on it by the magnetic field is given by

$$\mathbf{F} = \int d\mathbf{F} = I \int d\mathbf{l} \times \mathbf{B} = I(\mathbf{l} \times \mathbf{B}) \tag{5.17}$$

The magnitude of the force is

$$F = IlB \sin \theta \tag{5.18}$$

The direction of the force is given by Eq. (5.17) and is in the direction of $(\mathbf{l} \times \mathbf{B})$ where the vector **l** is taken along the direction of the current. Notice that the direction of the force is perpendicular to both **l** and **B**. The relationship between **F**, *I* and **B** can be described pictorially by Fleming's left-hand rule which reads "Hold the thumb and the first two fingers of your left hand mutually at right angles to each other (as shown in Fig. 5.13). Then, if the forefinger points in the direction of the field and the second finger in the direction of the current, the thumb will point in the direction of motion of the conductor (or force)".

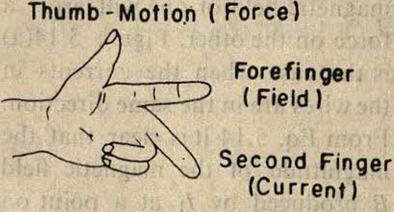


Fig. 5.13 *Fleming's left-hand rule*

Unit of B The unit of magnetic field *B* can be obtained from Eq. (5.14) which gives

$$B = \frac{F}{qv \sin \theta}$$

If

$$F = 1N$$

$$q = 1C$$

$$v = 1 \text{ ms}^{-1}$$

$$\theta = 90^\circ$$

then

$$B = \frac{1 \text{ N}}{1\text{C} \times 1 \text{ ms}^{-1} \times \sin 90^\circ}$$

$$= \frac{1 \text{ N}}{1 \text{ A} \times 1 \text{ m}} \qquad (\because \text{ C s}^{-1} = \text{A})$$

$$= 1 \text{ N A}^{-1} \text{ m}^{-1}$$

In the SI system, the unit of *B* is the tesla T. One tesla is one newton per ampere metre. *Magnetic field B is said to be 1 T if a* *charge of 1 C moving with a speed of 1 ms^{-1} at right angles to the field experiences a force of 1 N.*

A field of 1 tesla (T) is a very strong magnetic field. Very often, magnetic fields are expressed in terms of a smaller unit called the gauss (G).

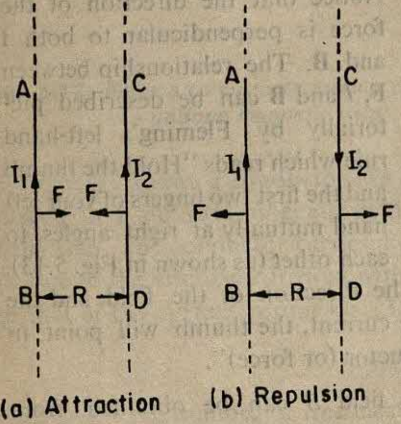
$$1\,G = 10^{-4}T$$

Force Between Two Parallel Wires Carrying Current

Consider two long parallel wires AB and CD separated by a certain distance R and carrying currents I_1 and I_2 as shown in Fig. 5.14. The magnetic field of one will exert a force on the other. Figure 5.14(a) is the case when the currents in the wires are in the same direction. From Eq. 5.14 it is clear that the magnitude of the magnetic field B produced by I_1 at a point on the other wire is given by

$$B = \frac{\mu_0 I_1}{2\pi R}$$

The direction of **B** is perpendicular to both I_1 and R, the perpendicular distance between the wires. **B** is therefore directed into the page and is at right angles to wire AB or (I_1). Since **B** is also perpendicular to I_2 the magnitude of force exerted by **B** on an element dl of wire CD is



(a) Attraction (b) Repulsion

Fig. 5.14 *Force between two parallel currents*

$$\mathrm{d}F = I_2\,\mathrm{d}lB \sin 90^\circ = I_2\,dlB$$

Substituting for B, we have

$$dF = \frac{\mu_0 I_1 I_2\,\mathrm{d}l}{2\pi R}$$

or

$$F = \frac{\mu_0 I_1 I_2 l}{2\pi R}$$

where l, the length of wire CD, is assumed to be large.

Force per unit length is

$$f = \frac{F}{l} = \frac{\mu_0 I_1 I_2}{2\pi R} \tag{5.19}$$

From Fleming's left-hand rule, this force on wire CD is to the left, i.e. towards the wire AB. Similar considerations apply for the force exerted on wire AB by the field of wire CD. Thus, we conclude that when the currents in the wires are in the same direction, the forces between them are attractive. On the other hand, if the currents are in opposite directions (anti-parallel), the two wires repel each other as shown in Fig. 5.14(b).

Definition of Ampere

The force between two infinitely long parallel wires carrying current has been used as the basis for the definition of a standard ampere of current in the SI system. From Eq. (5.19) it follows that if

$$I_1 = I_2 = 1 \text{ A}$$

$$R = 1 \text{ m}$$

then

$$f = \frac{\mu_0}{2\pi} = 2\times 10^{-7} \text{ Nm}^{-1}$$

for free space. Thus *the ampere is the strength of that steady current which, when flowing in two parallel infinitely long conductors of negligible cross-section placed in a vacuum at a distance of 1.0 m from each other, produces between the two conductors a force of 2× 10⁻⁷ newton per metre length.*

EXAMPLE 5.5 The battery of an automobile is connected to the starting motor by straight wires 70 cm long and 1.5 cm apart. If, at the time of starting, the current in the wires in 300 A, estimate the force between the wires. Is the force attractive or repulsive ?

Solution: The force per unit length between is the wires is

$$f = \frac{\mu_0 I_1 I_2}{2\pi R}$$

$$= \frac{4\pi\times 10^{-7}\times 300\times 300}{2\pi\times(1.5\times 10^{-2})} = 1.2 \text{ Nm}^{-1}$$

Therefore, the force between the wires ($l = 70$ cm $= 0.70$ m) is

$$F = fl$$

$$= 1.2\times 0.70 = 0.84 \text{ N}$$

This value is approximate since our formula for f holds strictly for infinitely long wires. Since the currents in the two wires are in opposite directions, the force is repulsive.

EXAMPLE 5.6 A straight horizontal conducting rod of length 0.45 m and mass 60 g is suspended by two vertical wires at its ends. A current of 5.0 A is set up in the rod through the wires. (a) What magnetic field should be set up normal to the conductor in order that the tension in the wires is zero? (b) What will be the total tension in the wires if the direction of current is reversed, keeping the magnetic field the same as before? (Ignore the mass of the wires, and assume $g = 9.8 \text{ ms}^{-2}$.)

Solution: (a) In order that the tension in the supporting wires is zero, the downward gravitational force mg on the rod must be balanced by an upward force BIl due to magnetic field, i.e.

$$BIl = mg$$

or

$$B = \frac{mg}{Il}$$

$$= \frac{(60 \times 10^{-3}) \times 9.8}{5.0 \times 0.45} = 0.26 \text{ T}$$

The direction of **B** is such that Fleming's left-hand rule gives an upward force.

(b) If the current is reversed, the direction of the force due to **B** becomes downwards, in the direction of the gravitational force. Hence the tension in the strings is

$$\begin{aligned} T &= BIl + mg \\ &= 0.26 \times 5.0 \times 0.45 + 0.06 \times 9.8 \\ &= 1.17 \text{ N} \end{aligned}$$

5.5 LORENTZ FORCE

The force experienced by a charge moving in space where both electric and magnetic fields exist is called the *Lorentz force*. Suppose there exists an electric field **E** at a certain point in space. The electric force $\mathbf{F}_e$ experienced by a charge q placed at that point is given by (see Chapter 1)

$$\mathbf{F}_e = q\mathbf{E}$$

An important property of the electric field is that the force experienced by the test charge placed at any point in the field is independent of whether the charge q itself is moving or not. Thus, if the electric field intensity at a point is **E**, the force experienced by a test charge q at rest or in motion is $q\mathbf{E}$.

In section 5.4, we have seen that the magnetic force $\mathbf{F}_m$ experienced by a charge q moving with a velocity $\mathbf{v}$ in a magnetic field $\mathbf{B}$ is given by

$$\mathbf{F}_m = q(\mathbf{v} \times \mathbf{B})$$

Hence the Lorentz force (i.e. total force) on a charge moving with velocity $\mathbf{v}$ in an electric field $\mathbf{E}$ and magnetic field $\mathbf{B}$ is

$$\mathbf{F} = \mathbf{F}_e + \mathbf{F}_m$$

or

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \tag{5.20}$$

In section 5.9, we will discuss some interesting consequences of Lorentz force.

Relative Magnitudes of Electric and Magnetic Forces

It is interesting to compare electric and magnetic forces under similar conditions. From Coulomb's law (see chapter 1), the magnitude of the electric force between two charges q_1 and q_2 separated by a distance r is given by

$$F_e = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r^2} \tag{5.21}$$

Equation (5.19) gives the force per unit length between two parallel infinitely long wires carrying currents I_1 and I_2. This relation is true only near the centres of the wires. The exact magnitude of the magnetic force between two parallel wires (carrying currents I_1 and I_2) of elements dl_1 and dl_2 separated by a distance r is

$$F_m = \frac{\mu_0}{4\pi} \cdot \frac{I_1 I_2}{r^2} \cdot dl_1\, dl_2 \tag{5.22}$$

We have seen above [Eq. (5.15)] that

$$I_1\, dl_1 = en_0(A_1\, dl_1)v_1 = q_1 v_1$$

and

$$I_2\, dl_2 = en_0(A_2\, dl_2)v_2 = q_2 v_2$$

since $en_0(A_1\, dl_1)$ = electronic charge × number of free electrons per unit volume × volume of element dl_1
= electronic charge × number of free electrons in element dl_1
= q_1, the charge in element

Similarly $en_0(A_2\, dl_2) = q_2$, the charge in element dl_2. Equation (5.22) then reads

$$F_m = \frac{\mu_0}{4\pi} \cdot \frac{q_1 q_2}{r^2} \cdot v_1 v_2 \tag{5.23}$$

It must be clearly understood that there is no electric force between two current-carrying wires because a conductor carrying a current remains electrically neutral since charge cannot accumulate anywhere in a conductor. The rate at which electrons enter one end of the conductor is equal to the rate at which they leave the other end. Thus sending a current through a conductor does not disturb its electrical neutrality.

Taking the ratio of (5.21) and (5.23) we have

$$\frac{F_m}{F_e} = (\mu_0\epsilon_0)v_1v_2$$

We will see in chapter 9 that the product $(\mu_0\epsilon_0) = 1/c^2$, where c is the speed of light in free space. Thus

$$\frac{F_m}{F_e} = \frac{v_1v_2}{c^2} \tag{5.24}$$

Typically, the electron drift velocities v_1 and v_2 are of the order of $10^{-5}\ \mathrm{ms^{-1}}$ and the value of $c = 3\times10^8\ \mathrm{ms^{-1}}$. Thus the ratio $F_m/F_e \simeq 10^{-27}$. In other words, under similar conditions, the magnetic forces are much smaller than electric forces. Why then are magnetic forces observed and not washed out by electric forces? The reasons are as follows. First, matter is electrically neutral to a very high degree of accuracy, especially, conductors in which, due to free-electron movement, the electrical neutrality cannot be easily disturbed. Secondly, in an electrical current of even a moderate amount, e.g. 1 A, the number of electrons drifting together is very large. Consequently, the small magnetic forces on individual electrons add up to sizeable amount, making it observable. It follows from Eq. (5.24) that for charges drifting with speeds comparable to the speed of light, magnetic and electric forces are of the same order of magnitude.

5.6 AMPERE'S CIRCUITAL LAW

In electrostatics (chapter 2) we have seen that the basic laws of electrostatics could be rephrased in the form of a law (Gauss's Law) which gives a quantitative relationship between the surface integral of the electric field over a closed surface and the charge inside it. Is there a similar relationship between magnetic field and current? Since a charge produces an electric field and a current (moving charge) produces a magnetic field (Oersted's discovery), it is reasonable to expect that a magnetic analogue of Gauss's law can be found. Ampere was the first person to discover the relation between magnetic field **B** and current I. He did not formulate it in terms of the field; this was done later by Clarke Maxwell. This relationship is called *Ampere's circuital law* (or simply Ampere's law) which may be stated as follows:

The loop or line circuit integral of the magnetic field along a closed curve is proportional to the current threading or passing through the closed circuit, i.e.

$$\oint \mathbf{B}\cdot d\mathbf{l} = \mu_0 I \tag{5.25}$$

where μ_0 is the permeability of free space.

The rigorous proof of the law is outside the scope of this book. We will, however, show that it is plausible by the following argument.

Consider an infinitely long straight wire carrying a current [Section 5.3, Fig. 5.3(a)]. We have seen that the magnitude of the magnetic field due to the current carrying wire at a distance r from it is given by [Eq. (5.7)]

$$B = \frac{\mu_0 I}{2\pi r} \tag{5.26}$$

The lines of force of the field are shown in Fig. 5.3(b); they are concentric circles in a plane perpendicular to the plane of the wire with the wire at the common centre. The magnetic field at all points on a circle has the same magnitude and its direction is perpendicular to the radius vector. Let us consider one such circle of radius r as shown in Fig. 5.15.

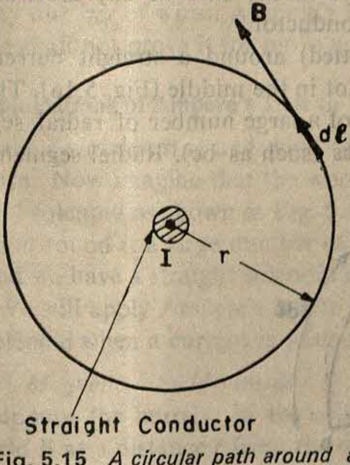


Fig. 5.15 *A circular path around a straight wire carrying a current I. The central dot indicates current I emerging from the page, the wire being perpendicular to the page*

We divide the circular path into small elements of length $d\mathbf{l}$ and evaluate the quantity $\mathbf{B}\cdot d\mathbf{l}$ and integrate it over the entire path. Notice that $\mathbf{B}$ and $d\mathbf{l}$ are both tangential to the circle at every point, i.e. the angle between them is zero so that the scalar product $\mathbf{B}\cdot d\mathbf{l} = B\,dl \cos 0^\circ = B\,dl$. Thus we have

$$\oint \mathbf{B}\cdot d\mathbf{l} = \oint B\,dl = B\oint dl$$

The magnitude B of magnetic field has been taken out of the sign of integration since B has a constant value at every point on the circle. Now $\oint dl$ is simply the circumference of the circle which is $2\pi r$. Hence

$$\oint \mathbf{B} \cdot d\mathbf{l} = B(2\pi r)$$

Substituting for B from Eq. (5.26) we have

$$\oint \mathbf{B} \cdot d\mathbf{l} = \mu_0 I$$

which is Ampere's law stated above. Although we have established the law for the special case of a field due to a straight and long current-carrying conductor, Ampere's law is true in general for any magnetic field (except for a magnetic field arising from a time-varying electric field, for which case the law has to be modified), for any assembly of currents and for any path of integration.

Generalization

We have established the correctness of Ampere's law by performing the integration along a circular path around a straight current conductor. We will now show that the law holds in the case of any arbitrary plane closed path around a straight conductor.

Consider a closed path (shown dotted) around a straight current carrying conductor shown by a big dot in the middle (Fig. 5.16). The path can be imagined to be made up of a large number of radial segments (such as ab) and circular arcs (such as bc). Radial segments

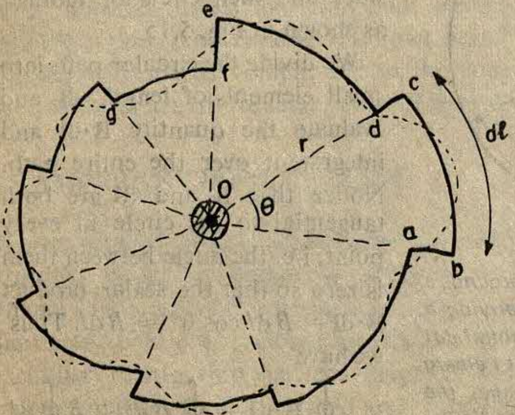


Fig. 5.16 *A closed loop of arbitrary shape around a straight current-carrying conductor placed perpendicular to the page*

(such as ab, cd, ef, . . ., etc.) and circular arcs (such as bc, de, fg, . . ., etc.) are shown in the figure. Since **B** is every where perpendicular to the radius, the contribution of all the radial segments to $\oint \mathbf{B}\cdot d\mathbf{l} = 0$. Therefore $\oint \mathbf{B}\cdot d\mathbf{l}$ is simply the value over all the circular segments. Let us evaluate the contribution of one such circular segment (such as bc) of arc length bc $= dl$ which is a part of a circle of radius, say, r as shown. The line integral of **B** along this segment is ($\because$ $dl = r\theta$)

$$\oint_{bc} \mathbf{B}\cdot d\mathbf{l} = \frac{\mu_0 I}{2\pi r}(r\theta) = \frac{\mu_0 I\theta}{2\pi}$$

where θ is the angle (in radians) subtended by the arc bc at the centre O. Notice that the contribution of the circular segment is independent of the radius r, and is proportional to the angle θ sustended by that arc at the centre. Since the angle θ for the entire closed loop is 2π, the integral over the entire loop is

$$\oint \mathbf{B}\cdot d\mathbf{l} = \frac{\mu_0 I}{2\pi}\times 2\pi = \mu_0 I$$

which is Ampere's law. This calculation is based on an infinitely long current-carrying wire. It can be shown that the law, in fact, is true for any number of wires. Ampere's law is thus completely general, though we shall not prove it for move complicated circuit shapes.

Applications of Ampere's Law to Toroidal and Straight Solenoids

Take a wooden ring and wind round it a large number of turns of a wire. Now imagine that the wooden ring is not there; we have a toroidal solenoid as shown in Fig. 5.17(a). Next take a round pencil and wind round it a large number of turns of a wire. Remove the pencil and we have a straight solenoid as shown in Fig. 5.17(b).

We will apply Ampere's law to find the magnetic field set up by the solenoid when a current is passed through it.

(a) *Magnetic Field Inside a Current-carrying Toroidal Solenoid* Suppose the current in the coils is clockwise and we wish to find the field **B** at a distance r from the centre [Fig. (5.17(a)]. From symmetry, the field has the same magnitude at all points on the circumference of the circle of radius r. The direction of the field is tangential to the circle at every point. Hence the line integral of **B** around the circular path of radius r [shown dashed in Fig. 5.17(a)] is

$$\oint \mathbf{B}\cdot d\mathbf{l} = B\oint dl = B(2\pi r) \tag{5.27}$$

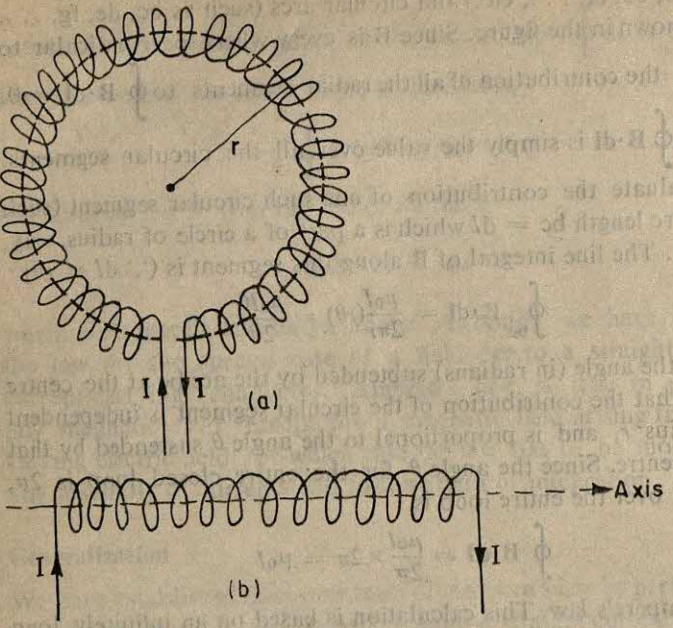


Fig. 5.17 (a) *Toroidal solenoid, and (b) straight solenoid*

From Ampere's circuital law,

$$\oint \mathbf{B}\cdot d\mathbf{l} = \mu_0 I$$

where I is the current threading the area bounded by the path of integration. Let i be the current in the turns of the solenoid. Notice that I is not the same as i because the path of integration (which is a circle of radius r) encloses more than one turn. If N is the total number of turns in the solenoid, then

$$I = Ni$$

so that

$$\oint \mathbf{B}\cdot d\mathbf{l} = \mu_0 Ni \tag{5.28}$$

From Eqs (5.27) and (5.28), we get

$$B(2\pi r) = \mu_0 Ni$$

or

$$B = \frac{\mu_0 Ni}{2\pi r} \tag{5.29}$$

If n is the number of turns *per unit length* of the toroid, then $n = N/2\pi r$. Hence in terms of n, Eq. (5.29) becomes

$$B = \mu_0 ni \tag{5.30}$$

Notice that B is independent of r, depending only on the current (i) and the number of turns per unit length (n). The field inside a toroidal solenoid is constant in magnitude and is always tangential. Toroidal solenoids are used to produce magnetic fields or as inductances or in transformers (see chapter 8).

(b) *Magnetic Field Inside a Long Current-carrying Straight Solenoid* A straight solenoid may also be visualized as follows. Imagine the radius of the toroidal solenoid becoming bigger. Any small portion of it will then appear straight. Since Eq. (5.30) is independent of the radius (r), it follows that the magnetic field inside a straight solenoid of n turns per unit length and carrying a current i is given by

$$B = \mu_0 ni \tag{5.31}$$

The direction of the field is parallel to the axis of the solenoid [Fig. 5.17(b)] and is fairly uniform except at the ends. Strictly speaking, Eq. (5.31) holds for an infinitely long solenoid. Such a solenoid is used if a uniform field is desired in a limited region.

EXAMPLE 5.7 A solenoid is 1.0 m long and 3.0 cm in mean diameter. It has five layers of windings of 800 turns each and carries a current of 5.0 A. Find the magnitude of the magnetic field at its centre.

Solution: The magnitude of magnetic field at or near the centre of a straight solenoid is [(Eq. (5.31)]

$$B = \mu_0 ni \tag{i}$$

where n is the number of turns per unit length. Note that we can use this expression for B even if the solenoid has more than one layer of windings because the diameter of the windings does not enter in the expression. Now, to find n, simply multiply the number of turns per layer by the number of layers and divide the product by the length of the solenoid. Thus

$$n = \frac{800 \times 5}{1.0} = 40{,}000 \text{ m}^{-1}$$

Now $i = 5.0$ A and $\mu_0 = 4\pi \times 10^{-7}$ TmA^{-1}. Using these values in (i) we get

$$B = 2.5 \times 10^{-2} \text{ T}$$

5.7 CURRENT LOOP AND MAGNETIC DIPOLE

A current-carrying straight solenoid behaves like a bar magnet. It will attract iron objects such as common pins. If suspended freely, it will orient itself in the north-south direction just as a bar magnet does. A bar magnet has two opposite poles at its ends; it is called a magnetic dipole. To see this connection, let us consider a circular current loop. In Section 5.3, we have seen that the magnitude of the magnetic field at a distance r from the centre of the loop and along the axis passing through it (in a direction perpendicular to the plane of the loop) is given by [see Eq. (5.11)]

$$B = \frac{\mu_0}{2} \cdot \frac{Ia^2}{(r^2 + a^2)^{3/2}}$$

where a is the radius of the circular loop. At distances $r \gg a$, this expression reduces to

$$B = \frac{\mu_0}{2} \frac{Ia^2}{r^3} = \left(\frac{\mu_0}{4\pi}\right) \frac{2I(\pi a^2)}{r^3}$$

or

$$B = \left(\frac{\mu_0}{4\pi}\right) \cdot \frac{2IA}{r^3} \qquad (5.31)$$

where $A = \pi a^2$ is the area of the circular loop. Let us compare it with the electric field due to an electric dipole at a point far from and at a point on the dipole axis. This field has been calculated in Section 1.12 and has a magnitude given by [see Eq. (1.17)]

$$E = \frac{1}{4\pi\epsilon_0} \cdot \frac{2p}{r^3} \qquad (5.32)$$

where p is the electric dipole moment. Notice that B due to a current loop and E due to an electric dipole have the same $(1/r^3)$ dependence on distance. Equation (5.31) may be written as

$$B = \left(\frac{\mu_0}{4\pi}\right) \cdot \frac{2m}{r^3} \qquad (5.33)$$

where $m = IA$. By analogy with electric dipole moment p in Eq. (5.32), quantity m is called the *magnetic dipole moment*; it is a vector quantity:

$$\mathbf{m} = I\mathbf{A} \qquad (5.34)$$

where **A** is the area vector.

Current loops, compass needles and bar magnets are all magnetic dipoles with north and south poles. One face of a current loop is the north pole and the other face the south pole.

5.8 TORQUE ON A CURRENT LOOP AND GALVANOMETER

We will now show that a current loop placed in a magnetic field experiences a torque. Comparing it with the torque experienced by an electric dipole in an electric field, we will again show that a current loop is a magnetic dipole. This dipole is used to measure currents (see the next sub-section on 'galvanometer').

Consider a rectangular loop PQRS of wire of sides a and b suspended in a uniform external magnetic field $\mathbf{B}$. Let θ be the angle between the direction of $\mathbf{B}$ and that of the vector perpendicular to the plane of the coil [see Fig. 5.18(a)]. Let a current I flow in the loop as shown.

The net force on the loop is the resultant of the forces on the four sides of the loop. Consider first the arms PQ and RS. It is easy to see that the force on each arm is the same. These equal forces are in opposite directions and are collinear (along the suspension shown by dashed line). Therefore, these forces give rise to no net force or torque and hence they do not have any effect on the motion of the coil.

Next consider forces on arms SP and QR. The force on each arm is $F = BIb$. The forces on SP and QR are in opposite directions but not collinear. The perpendicular distance between the two forces is $a \sin \theta$, as is obvious from Fig. 5.18(b). They, therefore, form a couple. The magnitude of the torque acting on the loop is

$$\begin{aligned} \tau &= \text{force} \times \text{perpendicular distance (or moment arm)} \\ &= BIb \times (a \sin \theta) \\ &= BIA \sin \theta \qquad (5.35(a)) \end{aligned}$$

or

$$\tau = mB \sin \theta \qquad (5.35(b))$$

where $m = IA$ is the magnitude of the magnetic dipole moment. In vector notation, the torque vector is given by

$$\boldsymbol{\tau} = \mathbf{m} \times \mathbf{B} \qquad (5.36)$$

This complete equation gives the magnitude as well as direction of $\boldsymbol{\tau}$. The direction of $\boldsymbol{\tau}$ is such that it rotates the loop in the clockwise sense about the axis of suspension. Expression (5.36) for torque on a current loop in a magnetic field is similar to expression (1.22) of Chapter 1 for torque on electric dipole in an electric field. Thus again we find that a current loop is a magnetic dipole.

If we have a rectangular coil of N turns, the torque multiplies N times and becomes

$$\tau = NIBA \sin \theta \qquad (5.37)$$

This equation can be shown to hold for all plane current loops of area A, irrespective of whether they are rectangular or not. The torque on a current loop is the operating principle of the electric motor and

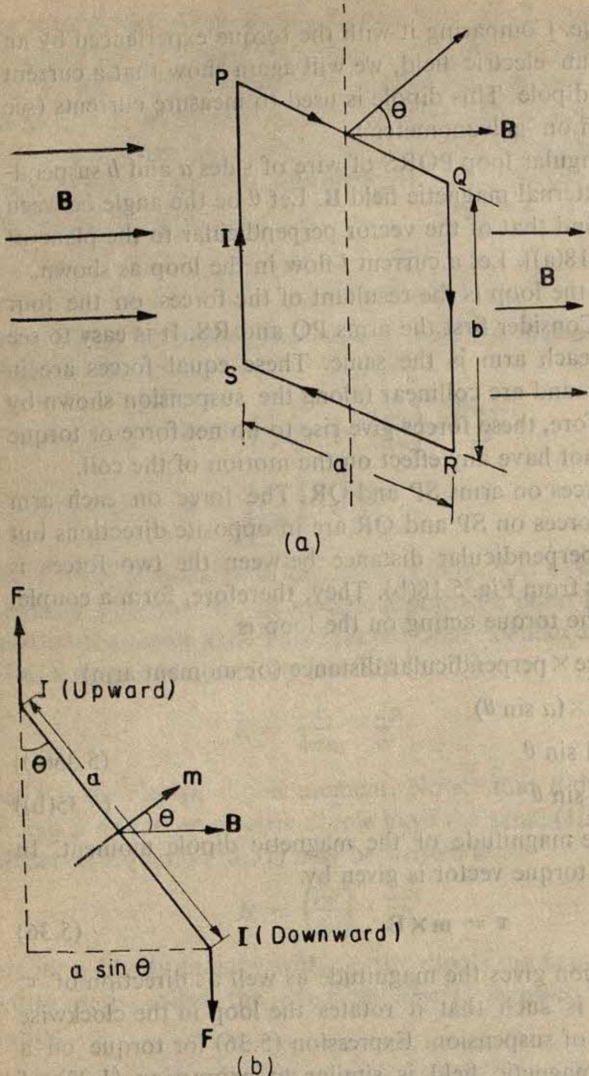


Fig. 5.18 *(a) A rectangular current loop in a uniform external magnetic field, and (b) top view of the loop with side SP coming out of the paper and side QR going in*

most electric meters used for measuring currents and voltages called galvanometers.

The Moving Coil Galvanometer

A galvanometer is a device for measuring electric currents.

Figure 5.19 illustrates the essential parts of moving coil galvanometer. It consists of a rectangular coil of many turns of an insulated wire suspended between the two pole pieces of a strong permanent horseshoe magnet. One end of the coil is attached to a fine phosphor bronze strip which: (i) serves as a suspension for the coil, and (ii) serves as one of the leads or terminals for passing current. The other end of the coil is attached to a loosely coiled strip which serves as the other lead as shown. The upper end of the suspension is attached to a torsion head which can be rotated to set the coil in the zero position. A mirror is fixed on the phosphor bronze strip by means of which the deflection of the coil is measured with a lamp-and-scale arrangement. To level the instrument, levelling screws (not shown in the figure) are provided at the base.

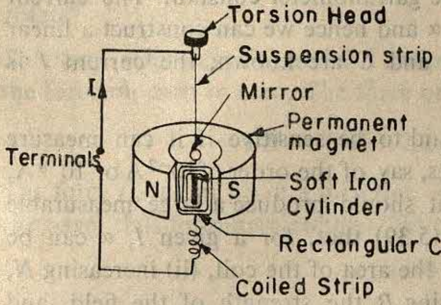


Fig. 5.19 *Essential parts of a moving coil galvanometer*

The working of the galvanometer is based on the fact that when a current is passed through a coil placed in a magnetic field, it experiences a torque given by Eq. (5.37). Under the influence of torque the coil rotates producing a twist in the suspension strip which provides the counter-torque that opposes and cancels out the magnetic torque τ resulting in a steady angular twist α corresponding to the current I. The counter torque is given by $k\alpha$ where k is the torque required to produce a unit angular twist in the suspension strip. When the coil comes to rest, then from Newton's third law, the action torque given by Eq. (5.37) and the reaction torque $k\alpha$ must be equal. Thus,

$$NIBA \sin \theta = k\alpha$$

or

$$I = \left(\frac{k}{NBA}\right) \cdot \frac{\alpha}{\sin \theta} \tag{5.38}$$

The relation between I and α is not linear because α and θ are related. Hence we have to construct a non-linear scale to measure currents and the instrument has to be calibrated using known currents. We can get rid of the factor sin θ in Eq. (5.38) if we use a radial magnetic field instead of the horizontal field. This is done by making the faces of the magnetic concave and inserting a soft-iron core (Fig. 5.20). The field lines are then radial and $\theta = 90°$ so that $\sin\theta = 1$ and Eq. (5.38) reduces

$$I = \left(\frac{k}{NBA}\right)\cdot\alpha = C\alpha \qquad (5.39)$$

where $C = (k/NBA)$ is called the galvanometer constant. The current now is directly proportional to α and hence we can construct a linear scale to read the deflection. If α and C are known, the current I is measured.

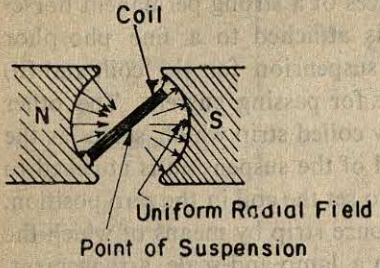


Fig. 5.20 *Radial magnetic field*

Sensitivity A galvanometer is said to be sensitive if it can measure or detect extremely small currents, say, of the order of 10^{-7} A or 10^{-8} A. In other words, a small current should produce a large, measurable deflection. It is clear from Eq. (5.39) that, for a given I, α can be increased by: (i) increasing A, the area of the coil, (ii) increasing N, the number of turns, (iii) increasing B, the strength of the field, and (iv) decreasing k, the restoring couple per unit twist of the suspension wire. It is for this reason that the suspension wire is made of phosphor bronze because its k is very low.

A soft iron cylinder is inserted inside the coil to concentrate the lines of force and to make the field radial. The concentration of lines of force increases the value of B and consequently the sensitivity of the galvanometer. To read deflection, a pointer may be attached to the suspension wire which may move over a scale. A more sensitive method is to attach a small mirror to the suspension wire and employ a lamp-and-scale arrangement as shown in Fig. 5.21.

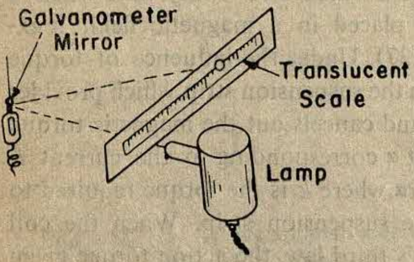


Fig. 5.21 *Lamp-and-scale arrangement*

EXAMPLE 5.8 A rectangular coil of 50 turns carries a current of 2.0 A and is placed in a magnetic field B of 0.25 T, as shown in Fig. 5.22.

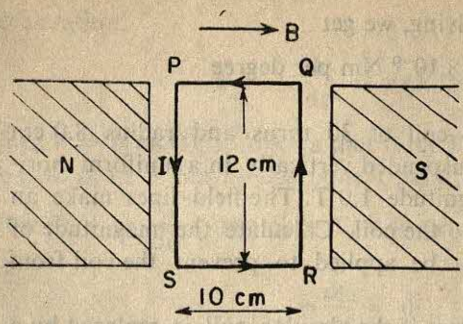


Fig. 5.22

Find the torque acting on the coil. In what direction will the coil rotate?

Solution: The sides PQ and RS are parallel to the field lines; hence the force on each in zero. The force on vertical sides PS and QR is

$$f = BIl = 0.25 \times 2.0 \times 0.12 = 0.06 \text{ N}$$

The force f acts on PS in a direction out of the page and on QR into the page, thus forming a couple. The total force is

$$F = 50 \times 0.06 = 3.0 \text{ N}$$

$$\therefore \quad \text{Torque } \tau = 3.0 \text{ N} \times 0.10 \text{ m}$$

$$= 0.3 \text{ N m}$$

The torque tends to rotate the side PS out of the page and QR into the page.

EXAMPLE 5.9 A rectangular coil of area 5.0×10^{-4} m² and 60 turns is pivoted about one of its vertical sides. The coil is in a radial horizontal field of 90 G. What is the torsional constant of the hairsprings connected to the coil if a current of 0.20 mA produces an angular deflection of 18°.

Solution:

$$B = 90 \text{ G} = 90 \times 10^{-4} \text{ T}, \qquad A = 5.0 \times 10^{-4} \text{ m}^2$$

$$\theta = 90^\circ \text{ (for radial field)}, \qquad N = 60$$

$$I = 0.20 \text{ mA} = 0.20 \times 10^{-3} \text{ A}, \ \alpha = 18^\circ$$

From Eq. (5.39), the torsional constant of the hairsprings is given by

$$k = \frac{INBA}{\alpha}$$

Using the above values and solving, we get

$$k = 3.0 \times 10^{-9} \text{ Nm per degree}$$

EXAMPLE 5.10 (a) A circular coil of 30 turns and radius 8.0 cm carrying a current of 6.0A is suspended vertically in a uniform horizontal magnetic field of magnitude 1.0 T. The field lines make an angle of 60° with the normal to the coil. Calculate the magnitude of the counter torque that must be applied to prevent the coil from turning.

(b) Would your answer change if the circular coil is replaced by a coil of some irregular shape that encloses the same area? (All other particulars are also unaltered.)

Solution: (a) Magnitude of the counter torque = magnitude of the torque due to **B**

$$= NIBA \sin\theta$$

$$= 30 \times 6.0 \times 1.0 \times (\pi \times 0.08 \times 0.08) \times \sin 60^\circ$$

$$= 3.1 \text{ Nm}$$

(b) The answer would not change since the torque on a planar coil in a field $\boldsymbol{\tau} = NI\mathbf{A}\times\mathbf{B}$ is true for a planar loop of any shape.

5.9 MOTION OF A CHARGED PARTICLE IN A UNIFORM MAGNETIC FIELD

Forces acting on charged particles moving in magnetic and electric fields are used to deflect, focus and accelerate beams of charged particles in many practical devices such as the TV tube (which is just a cathode ray tube) and the cyclotron. In this section we will discuss the motion of a charge in a uniform magnetic field. In the next section we will deal with the motion of a charge when an electric field is also present.

Suppose a particle, carrying a charge q and having a velocity **v** enters a uniform magnetic field **B** directed, say, along the z-axis of a rectangular coordinate system. What path or trajectory does the charged particle follow in the field? Newton's second law of motion provides the answer to this question. We know that the charged particle experiences magnetic Lorentz force

$$\mathbf{F} = q(\mathbf{v}\times\mathbf{B})$$

which imparts an acceleration $d\mathbf{v}/dt$ to the particle. If m is the mass of the particle, we have from Newton's second law (force = mass ×

acceleration)

$$\mathbf{F} = m\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = q(\mathbf{v}\times\mathbf{B}) \tag{5.40}$$

Since $\mathbf{F} = q(\mathbf{v}\times\mathbf{B})$, it follows that force **F** is perpendicular to magnetic field **B**, i.e. **F** has no component parallel to **B**. Thus $F_{\parallel}(=F_z) = 0$. Now $F_z = m\frac{\mathrm{d}v_z}{\mathrm{d}t}\cdot$ Since $F_z = 0$, it is clear that

$$m\frac{\mathrm{d}v_z}{\mathrm{d}t} = 0$$

Hence $v_z(=v_{\parallel})$ = constant, i.e. the component of the velocity in the direction of the magnetic field cannot change with time. The kinetic energy $\frac{1}{2}m(v_{\parallel}^2)$ associated with the parallel component of velocity is also constant.

Furthermore, the magnetic force $\mathbf{F} = q(\mathbf{v}\times\mathbf{B})$ is perpendicular also to velocity **v** (or the direction of motion) of the particle. Hence it can do no work on the particle. Hence *no work is done on a charged particle moving in a uniform magnetic field.* Remember, this result holds only if the magnetic field is uniform, i.e. it is constant in space and time. This implies that the kinetic energy of the particle must remain constant, i.e.

$$\mathrm{KE} = \frac{1}{2}mv^2 = \text{constant}$$

This can be explicitly proved as follows:

The rate of change of kinetic energy is

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}(\mathrm{KE}) &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}mv^2\right) \\
&= \frac{1}{2}m\frac{\mathrm{d}}{\mathrm{d}t}(v^2) && (\because m = \text{constant}) \\
&= \frac{1}{2}m\frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{v}\cdot\mathbf{v}) && (\because v^2 = \mathbf{v}\cdot\mathbf{v}) \\
&= \frac{1}{2}m\left(\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t}\cdot\mathbf{v} + \mathbf{v}\cdot\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t}\right) \\
&= \left(m\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t}\right)\cdot\mathbf{v} && \left(\because \frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t}\cdot\mathbf{v} = \mathbf{v}\cdot\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t}\right) \\
&= \mathbf{F}\cdot\mathbf{v} && \left(\because \mathbf{F} = m\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t}\right)
\end{aligned}$$

or
$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}\,mv^2\right) = \mathbf{F}\cdot\mathbf{v}$$

But $(\mathbf{F}\cdot\mathbf{v}) = 0$ since $\mathbf{F}$ is perpendicular to $\mathbf{v}$ and the dot product of two perpendicular vectors is zero. Hence the total kinetic energy $\frac{1}{2}\,mv^2$ does not change with time, i.e.

$$\frac{1}{2}\,mv^2 = \frac{1}{2}\,m(v_x^2 + v_y^2 + v_z^2) = \text{constant}$$

or
$$\frac{1}{2}\,mv^2 = \frac{1}{2}\,m\,(v_\perp^2 + v_\parallel^2) = \text{constant}$$

where $v_\perp^2 = v_x^2 + v_y^2$ and $v_\parallel = v_z$. We have already seen above that

$$\frac{1}{2}\,m\,v_\parallel^2 = \text{constant}$$

Hence

$$\frac{1}{2}\,m\,v_\perp^2 = \text{constant} \tag{5.41}$$

Since $v_\perp^2 = v_x^2 + v_y^2$, it is clear that v_x and v_y are the perpendicular components of a resultant velocity $v_\perp$ as shown in Fig. 5.23 where $v_1^2 = v_x^2 + v_y^2$. Thus v_x and v_y are the perpendicular components of a resultant velocity v_1 as shown in Fig. 5.23.

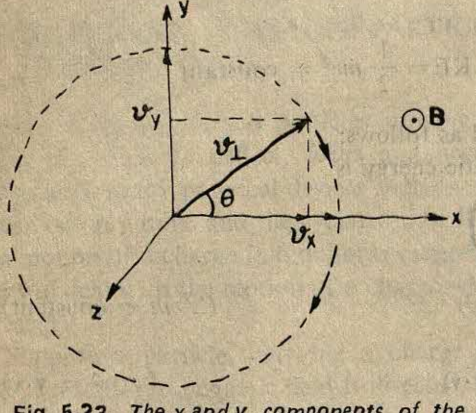


Fig. 5.23 *The x and y components of the velocity of a charged particle moving in a magnetic field directed out of the page*

Therefore, v_x and v_y are expressed in terms of $v_\perp$ as

$$v_x = v_\perp \cos\theta \tag{5.42a}$$

$$v_y = v_\perp \sin\theta \tag{5.42b}$$

where θ is angle which the resultant $v_\perp$ subtends with x-axis. Notice that, although v_x and v_y both change with time, $v_\perp$ [see Eq. (5.41)] cannot change with time.

Now let us consider the x and y components of Eq. (5.40), which are:

$$m\frac{\mathrm{d}v_x}{\mathrm{d}t} = qBv_y \tag{5.43a}$$

and

$$m\frac{\mathrm{d}v_y}{\mathrm{d}t} = -qBv_x \tag{5.43b}$$

Using Eqs (5.42a) and (5.42b) in Eq. (5.43a) we get (since $v_\perp$ = constant)

$$mv_\perp \frac{\mathrm{d}}{\mathrm{d}t}(\cos\theta) = qBv_\perp \sin\theta$$

or

$$-m\sin\theta\frac{\mathrm{d}\theta}{\mathrm{d}t} = qB\sin\theta$$

or

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = -\frac{qB}{m} \tag{5.44}$$

i.e angle θ changes at a constant rate. The arrowhead of vector $\mathbf{v}_\perp$, therefore, rotates with a constant angular velocity $\omega = \frac{\mathrm{d}\theta}{\mathrm{d}t}$ given by

$$\omega = \frac{qB}{m} \tag{5.45}$$

If q is negative and **B** directed out of the page along $+z$ axis, the sense of rotation will be clockwise as shown in Fig. 5.23. Integrating Eq. (5.44) we have

$$\theta = -\frac{qBt}{m} + \theta_0 = -\omega t + \theta_0 \tag{5.46}$$

where θ_0 is the constant of integration. Therefore Eqs (5.42 a and b) become

$$v_x = \frac{\mathrm{d}x}{\mathrm{d}t} = v_\perp \cos(-\omega t + \theta_0)$$

$$v_y = \frac{\mathrm{d}y}{\mathrm{d}t} = v_\perp \sin(-\omega t + \theta_0)$$

Integrating the above equations, we have

$$x = -\left(\frac{v_\perp}{\omega}\right)\sin(-\omega t + \theta_0) + x_0$$

$$y = +\left(\frac{v_\perp}{\omega}\right)\cos(-\omega t + \theta_0) + y_0$$

These equations give

$$(x - x_0)^2 + (y - y_0)^2 = \left(\frac{v_\perp}{\omega}\right)^2 \tag{5.47}$$

This is the equation of a circle with centre at (x_0, y_0) and radius $r = v_\perp/\omega$. Hence the trajectory of the particle is circular. The charged particle moves in a circle in the x-y plane of radius $r = v_\perp/\omega$ and angular velocity $\omega = qB/m$. This can also be understood as follows:

We have seen that the force acting on a moving charge in a uniform magnetic field: (i) is perpendicular to its velocity and (ii) cannot change the kinetic energy of the particle, i.e. the force cannot change the *magnitude* $v_\perp$ of the velocity $\mathbf{v}_\perp$, it can only change the direction of $\mathbf{v}_\perp$. These are exactly the conditions of uniform *circular motion*, as we have seen in Chapter 4 of the Class XI Physics textbook. Thus the charged particle follows a circular path and experiences a centrepetal force

$$F_\perp = \frac{mv_\perp^2}{r} \tag{5.48}$$

where r is the radius of the circular path. Now we know that

$$\mathbf{F} = q(\mathbf{v} \times \mathbf{B})$$

$$F_x = qv_yB$$

and

$$F_y = -qv_xB$$

$\therefore$

$$F_\perp^2 = F_x^2 + F_y^2 = q^2B^2(v_x^2 + v_y^2) = q^2B^2v_\perp^2$$

or

$$F_\perp = qBv_\perp \tag{5.49}$$

Equating (5.48) and (5.49) we get

$$qBv_\perp = \frac{mv_\perp^2}{r}$$

or

$$r = \frac{mv_\perp}{qB} \tag{5.50}$$

which gives the radius of the path. The angular velocity ω is given by

$$\omega = \frac{v_\perp}{r} = \frac{qB}{m}$$

which agrees with the expression (5.45) obtained above. The frequency ν measured in hertz is given by

$$\nu = \frac{\omega}{2\pi} = \frac{qB}{2\pi m} \tag{5.51}$$

Notice that the frequency of rotation is independent of the speed of the particle and the radius of its circular path; it depends only on its mass, charge and magnetic field. Equation (5.50) suggests that faster particles move in larger circles and slower particles move in smaller circles; but all particles take the same time to complete one revolution in the field. The frequency ν is a characteristic frequency of the charged particle in the magnetic field and is sometimes called the *cyclotron frequency* of the particle in the field because the particles circulate at this frequency in the cyclotron, which is a device used to accelerate charged particles.

We have seen that a charged particle moving in a uniform magnetic field has two kinds motion: a linear motion in the direction of the magnetic field (the z-direction) and a circular motion in a plane perpendicular to **B** which is the x-y plane. Hence the complete trajectory of the particle is a helix as shown in Fig. 5.24.

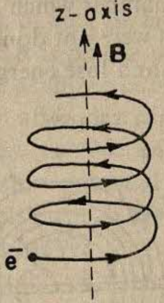


Fig. 5.24 *Actual trajectory of a negatively charged particle in a uniform magnetic field*

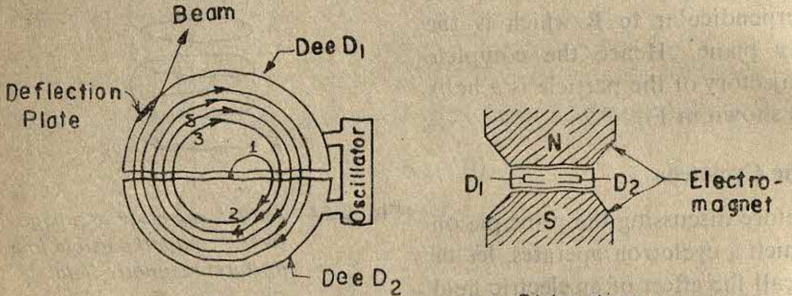
The Cyclotron

Before discussing the principle on which a cyclotron operates, let us recall the effect of an electric field **E** and a magnetic field **B** at a point, say P, in space on the motion of a charged particle at that point P. The electric field **E** exerts a force $q\mathbf{E}$ on the particle of charge q and this force is independent of the velocity of the particle. The particle is accelerated in the direction of the field, the magnitude of its acceleration being qE/m. On the other hand, the magnetic field, at right angles to the direction of motion of the particle throws the particle in a circular orbit.

The cyclotron is a device which is used to accelerate charged particles, such as protons, deuterons, α-particles, etc., to very high energies. The first cyclotron was designed and put into operation by Ernest Lawrence in 1932 in California, U.S.A.

Principle The principle on which a cyclotron works is based on the fact that an electric field can accelerate a charged particle and a magnetic field can throw it into a circular orbit in which the particle revolves with a frequency that does not depend on its speed. A modest potential difference is used as a source of electric field. If a charged particle is made to pass through this potential difference a number of times, it will acquire an enormously large velocity.

Construction A cyclotron consists of a hollow cylinder made from a copper sheet. The cylinder is split into two parts, D_1 and D_2. Each part is known as a dee. Figure 5.25(a) shows the top view and Fig. 5.25(b) the side view of a cyclotron. At the centre of the dees is the source of positive ions. The dees are kept between the pole pieces of a strong electromagnet whose field **B** points out of the plane of Fig. 5.25(a). A high frequency alternating voltage is applied between the dees by an electric oscillator whose frequency can be adjusted. The strength of the magnetic field is of the order of 1–2 T. The magnitude (peak value) of the electric field is about 10^4 to 10^5 V and its frequency is about 10^6 to 10^7 Hz. The whole assembly is housed in a vacuum chamber which is evacuated to a pressure of about 10^{-6} mm Hg. If this were not done, the ions would continually collide with air molecules and lose energy.



(a) Top view (b) Side View

Fig. 5.25 *The principle of a cyclotron*

The ion source at the centre of the cyclotron consists of deuterium which is bombarded with high energy electrons. This bombardment produces many positive ions. Many of these ions are free deuterons which emerge from a hole in the ion source and are available to be accelerated.

Working Suppose a positive ion, emerging from the ion source, finds that it is facing a dee, say D_1, which is negative at that instant of time. It will accelerate toward this dee and will enter it. Once the particle enters the dee, it is screened from the electrical forces by the metallic walls of the dee. However, the magnetic field is not screened by the dees. Hence the ion bends in a circular path whose radius, which depends on the velocity, is given by

$$r = \frac{mv_\perp}{qB}$$

After describing a semicircle, marked 1, the particle emerges from dee D_1 and finds itself in an electric field. Let t_0 be the time taken by it to describe the semicircle 1.

$$t_0 = \frac{T}{2} = \frac{1}{2\nu} = \frac{\pi m}{qB}$$

Let us assume that during this time, the accelerating potential has changed sign, i.e. dee D_2 is now negative and D_1 is positive. Thus, the ion again faces a negative dee (D_2) and is further accelerated and again describes a semicircle marked 2 but this time in dee D_2. Since the speed of the particle is increased, the radius of semicircle 2 is larger than that of semicircle 1. Since the time period T is independent of velocity, the time of passage through dee D_2 is still t_0. This process is repeated. The particle describes semicircles 3, 4, 5, etc., of larger radii until it reaches the outer edge of the dee where it is pulled out of the cyclotron by a negatively charged deflector plate as shown in Fig. 5.25a. Thus the particle acquires additional energy after describing each semicircle. After a number of such stages, the particle energy can become enormously large. With an accelerating potential of 10^5 V the energy of the particle after 100 such passages through the field can become as high as 10 million electron volts (10 MeV).

The key to the operation of cyclotron is that the characteristic frequency at which a charged particle circulates in a magnetic field must be equal to the fixed frequency of the electric oscillator, or

$$\nu = \nu_0$$

This is the condition of *resonance* which states that, if the energy of a particle is to be increased, then energy from the oscillator must be fed to it at a frequency ν_0 that is equal to the natural frequency at which the particle circulates in the field.

From Eq. (5.51), we can write the resonance condition as

$$\frac{qB}{2\pi m} = \nu_0 \tag{5.52}$$

For a given ion q/m is fixed. For a given B, the frequency ν_0 of the oscillator is adjusted until Eq. (5.52) is satisfied. The oscillator is then said to be 'tuned' to the field. Usually ν_0 is fixed and B is varied to meet the requirement in Eq. (5.52).

The energy of the particle emitted from the deflector plate depends on the radius R of the dees. From Eq. (5.52) the velocity of the particle circulating at the radius R is given by

$$v_\perp = \frac{qBR}{m} \tag{5.53}$$

The kinetic energy of the particle emerging from the cyclotron is

$$K = \frac{1}{2} mv_\perp^2 = \frac{q^2B^2R^2}{2m} \qquad (5.54)$$

Uses The high energy particles produced in a cyclotron are used for bombarding nuclei or for producing other high energy particles, such as neutrons, by collisions. These fast neutrons are used in atomic reactors. These high energy particles act as powerful bullets which can smash atoms. They have been used in the artificial transmutation of elements and also in causing induced radioactivity in various elements.

EXAMPLE 5.11 A uniform magnetic field of magnitude 1.0 T points horizontally from south to north. A 5.0 MeV proton moves vertically downward through this field. (i) What is the magnitude of the force acting on the proton and (ii) What is its direction? The mass of proton $= 1.7 \times 10^{-27}$ kg.

Solution: (i) The kinetic energy of the proton is

$$\begin{aligned} K &= 5.0 \text{ MeV} \\ &= 5.0 \times 10^6 \text{ eV} \\ &= 8.0 \times 10^{-13} \text{ J} \end{aligned}$$

Its speed v is given by the relation $K = \frac{1}{2} mv^2$, giving

$$\begin{aligned} v &= \left(\frac{2K}{m}\right)^{1/2} \\ &= \left(\frac{2 \times 8.0 \times 10^{-13}}{1.7 \times 10^{-27}}\right)^{1/2} \\ &= 3.1 \times 10^7 \text{ ms}^{-1} \end{aligned}$$

Now the equation $F = qvB \sin\theta$ gives the force, i.e.

$$F = 1.6 \times 10^{-19} \times 3.1 \times 10^7 \times 1.0 \times \sin 90^\circ = 4.96 \times 10^{-12} \text{ N}$$

The proton also experiences a gravitational force $= mg$

$$= 1.7 \times 10^{-27} \times 9.8 = 16.7 \times 10^{-27} \text{ N}$$

We notice that the force in a magnetic field is about 3×10^{14} times stronger than the gravitational force.

(ii) The relation $\mathbf{F} = q(\mathbf{v} \times \mathbf{B})$ tells us that the direction of the deflecting force is to the east. If the particle had a negative charge, it would experience a force to the west. This follows if we replace q by $-q$.

EXAMPLE 5.12 An electron of energy 10,000 eV describes a circular path in a plane at right angles to a uniform magnetic field of 0.01 T. (i) What is the radius of the circular orbit? (ii) What is the cyclotron frequency? (iii) What is the period of its revolution? (iv) What is the direction of revolution as viewed by an observer looking in the direction of the field?

Solution: (i) The velocity of the electron of energy K is found from the relation

$$v_{\perp} = \sqrt{\frac{2K}{m}} \tag{i}$$

Here

$$K = 10{,}000 \text{ eV} = 1.6 \times 10^{-15} \text{ J}$$

$$m = 9.1 \times 10^{-31} \text{ kg}$$

Using these values in (i) gives

$$v_{\perp} = 5.9 \times 10^{7} \text{ ms}^{-1}$$

The radius of the orbit is given by

$$r = \frac{mv_{\perp}}{qB}$$

with $B = 0.01$ T, we have

$$r = 3.4 \times 10^{-2} \text{ m}$$

(ii) The cyclotron frequency is

$$\nu = \frac{qB}{2\pi m} = 2.8 \times 10^{8} \text{ Hz}$$

(iii) The time period of revolution is

$$T = \frac{1}{\nu} = 3.6 \times 10^{-9} \text{ s}$$

(d) If the charge moving to the right perpendicular to **B** were negative, the force **F** would point downward. Thus, an electron revolves in a clockwise sense as seen by observer looking in the direction of the field **B**. A positive charge would revolve in the anti-clockwise sense when viewed along the field.

EXAMPLE 5.13 Deuterons are accelerated in a cyclotron that has an oscillatory frequency of 10^7 Hz and a dee radius of 50 cm. (i) What is the strength of magnetic field needed to accelerate the deuterons? (ii) What is the energy of deuterons emerging from the cyclotron?

Mass of a deuteron $= 3.34\times10^{-27}$ kg and charge of a deuteron $= 1.6\times10^{-19}$ C.

Solution: (i) We have

$$B = \frac{2\pi m\nu_0}{q}$$

$$= \frac{2\times3.14\times3.34\times10^{-27}\times10^7}{1.6\times10^{-19}} = 1.3 \text{ T}$$

(ii) The energy of the deuteron is [see Eq. (5.54)]

$$K = \frac{q^2B^2R^2}{2m}$$

Here R = radius of the dee = 50 cm = 0.5 m. Thus,

$$K = 1.6\times10^{-12} \text{ J}$$

$$= 10^7 \text{ eV} = 10 \text{ MeV}$$

5.10 MOTION OF A CHARGED PARTICLE IN PERPENDICULAR MAGNETIC AND ELECTRIC FIELDS

Let us consider a charged particle, for example an electron, moving with velocity **v** in a region where a uniform electric field **E** and a uniform magnetic field **B** are both present. The total Lorentz force on the electron is [setting $q = -e$ in Eq. (5.20) where $e = 1.6\times10^{-19}$C]

$$F = -e(\mathbf{E} + \mathbf{v}\times\mathbf{B}) \tag{5.55}$$

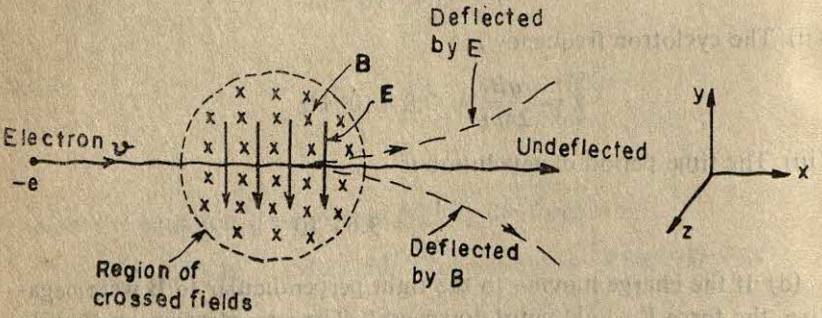


Fig. 5.26 *An electron passing through crossed electric and magnetic fields*

Now if **v** and **B** are mutually perpendicular, as shown in Fig. 5.26, (**v** is along positive x-direction and **B** is along the z-direction pointing into the page) and if the plane containing them (i.e. the x-z plane) is perpendicular to **E** (which is along negative y-direction), then the force ($-e\mathbf{E}$) will be along positive y-direction and the force $-e(\mathbf{v}\times\mathbf{B})$ along negative y-direction, i.e. the electric and magnetic forces will be collinear and in opposite directions. If the magnitudes of **E** and **B** are such that the two opposite forces are equal in magnitude, then the electron will go through the so-called 'crossed' fields undeflected.

Equation (5.55) tells us that the net force $\mathbf{F} = 0$ if

$$\mathbf{E} = -(\mathbf{v} \times \mathbf{B})$$

For given **E** and **B** fields, this will happen for a particular speed v given by

$$E = vB \sin\theta = vB \qquad (\because \theta = 90°)$$

or
$$v = \frac{E}{B} \tag{5.56}$$

where E and B are the magnitudes of **E** and **B**; their directions are shown in Fig. 5.26. This condition can be used to 'select' charged particles of a particular velocity out of a beam in which the particles are moving with different speeds. This method was used by J J Thomson to measure the charge to mass ratio (e/m) of an electron.

J J Thomson's Method for Determining e/m of an Electron

In 1897, J J Thomson measured the ratio of the charge e of an electron to its mass m by observing its deflection in a combined electric and magnetic field.

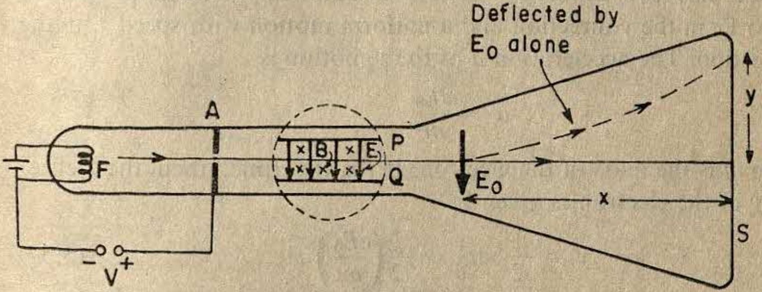


Fig. 5.27 *J J Thomson's method for determining e/m of an electron*

A modern version of Thomson's experiment is illustrated in Fig. 5.27. Electrons are emitted from a hot filament F and accelerated by applying a voltage V between the filament F (cathode) and anode A. The narrow beam of electrons emerging from the hole in A then enters the 'velocity selector' which is a region of crossed electric and magnetic fields. S is a fluorescent screen at the other end of the glass tube called the cathode ray tube. The tube is highly evacuated so that collisions with air molecules can be prevented. P and Q are two parallel plates across which a suitable potential difference is applied, thus producing an electric field (shown by arrows) between P and Q at right angles to the plates and in the plane of the paper. A magnetic field is applied (by placing the tube between the pole pieces of an electromagnet) in the same space (shown by the broken circle) so that the field is directed into the page (shown by crosses). Thus the electric and magnetic fields are mutually perpendicular to each other.

A study of Fig. 5.27 reveals that the electric field deflects the electrons upward and the magnetic field deflects them downward. The deflection of the beam can be observed by the deflection of the spot on the screen S. The fields **E** and **B** are adjusted until the upward deflection by the electric field is completely cancelled by the downward deflection by the magnetic field. The beam of electrons will then continue straight without any deflection. This happens when the force experienced by the electrons in the electric field is equal to that in the magnetic field, i.e.

$$v = \frac{E}{B}$$

This equation determines the velocity of electrons which travel undeflected in the crossed fields. Consider an electron beam with a velocity **v** selected in this manner. Suppose it is now deflected by a uniform electric field $\mathbf{E}_0$ alone (perpendicular to **v** as shown in the figure). The electron has two kinds of motion: a uniformly accelerated motion due to $\mathbf{E}_0$ in the y-direction and a uniform motion with speed v in the x-direction. The acceleration a of the y-motion is

$$a = \frac{eE_0}{m}$$

where m is the mass of the electron. If t is the time, then the deflection y of the electron is given by

$$y = \frac{1}{2}\,at^2 = \frac{1}{2}\left(\frac{eE_0}{m}\right)t^2 \qquad (5.57)$$

The distance x travelled in time t along the x-direction due to a uniform speed v is

$$x = vt \qquad (5.58)$$

Eliminating t from Eqs (5.57) and (5.58) we get

$$y = \left(\frac{eE_0}{2mv^2}\right)x^2$$

Since $y \propto x^2$, the trajectory of the electron is parabolic. Also

$$\frac{e}{m} = \frac{2v^2y}{E_0x^2} \qquad (5.59)$$

The speed v is determined from condition (5.58) by finding E and B fields which produce zero deflection of the spot on the screen. Having found v, we measure y and x for a known electric field E_0. Thus knowing v, x, y and E_0, the value of e/m is determined from Eq. (5.59). Thomson found that $e/m = 1.7 \times 10^{11}$ C kg^{-1}. The best present value is

$$\frac{e}{m} = (1.758802 \pm 0.000005) \times 10^{11} \text{ C kg}^{-1}$$

5.11 A SOLID IN PERPENDICULAR ELECTRIC AND MAGNETIC FIELDS: THE HALL EFFECT

In 1879, E.H. Hall, an American scientist, devised an experiment which not only decides the sign of the charge carriers but also determines the number of mobile charges in a conductor.

Consider a metal in the form of a flat strip (Fig. 5.28). A steady voltage is applied between the two faces R and S such that the resulting electric field **E** points downwards. The negative charges, therefore, drift upwards with a velocity $\mathbf{v}_d$, resulting in a current I flowing downwards (remember, conventionally, the direction of I is the direction in which the positive charges would drift).

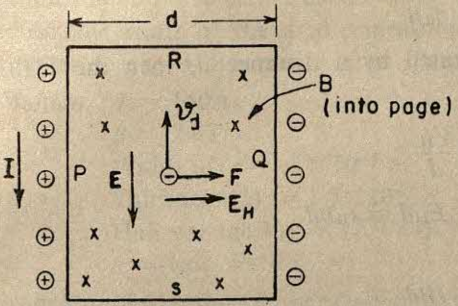


Fig. 5.28 *A metal strip placed in perpendicular electric and magnetic fields*

Now suppose a magnetic field **B** is applied perpendicular to the plane of the strip and directed into the page. Hall discovered that, due to **B**, a potential difference V develops between the other two faces P and Q of the strip. This is the Hall effect, which can be understood as follows:

We have seen in Chapter 3 that the current I and drift velocity are related as

$$\mathbf{I} = neA\,\mathbf{v}_d$$

where n = number of drifting charges per unit value, e = magnitude of charge and A the area of the strip. Thus

$$\mathbf{v}_d = \frac{\mathbf{I}}{neA} \tag{5.60}$$

Now, if the drifting charges are in a magnetic field **B** such that $\mathbf{v}_d$ is perpendicular to **B**, they experience a magnetic force **F** given by (see Fig. 5.28).

$$\mathbf{F} = \mathbf{v}_d \times \mathbf{B} = \frac{1}{neA}(\mathbf{I} \times \mathbf{B})$$

Because of this force, the charges tend to drift to the right toward the face Q of the strip. This drift is due to the magnetic field. Remember the charges (assumed negative) are also drifting upward toward the face R under an electric field **E**. Due to force **F** the negative charges will accumulate at face Q, leaving an equal positive charge at face P. Thus a potential difference V_H (called the *Hall potential difference*) is developed between faces P and Q. This produces an electric field $\mathbf{E}_H$ such that the force $e\mathbf{E}_H$ due to $\mathbf{E}_H$ always opposes the force $e(\mathbf{v}_d \times \mathbf{B}) = ev_dB \sin 90^\circ = ev_dB$ due to the magnetic field. The accumulation of charges stops when equilibrium is reached. i.e. when

$$eE_H = ev_dB$$

or

$$E_H = v_dB$$

If the faces P and Q are separated by a distance d, then the Hall voltage V_H is given by

$$E_H = \frac{V_H}{d}$$

or

$$V_H = E_Hd = v_dBd$$

which, using Eq. (5.60) reads

$$V_H = \frac{IBd}{neA} \tag{5.61}$$

In writing these equations we have ignored the signs of $\mathbf{v}_d$, $\mathbf{E}_H$ and **B** which are mutually perpendicular. We have assumed that the charge carriers in metals are electrons. For positive charge carriers, the sign of V_H would be negative. The sign of V_H can be ascertained by an actual measurement in such an experiment. Hall found that V_H was positive, showing that the charge carriers in metals are electrons. Measuring V_H and knowing I, B, d and A, we can determine the number density (n) of the moving charges from Eq. (5.61). The experimental values of n agree fairly well with theoretical values for monovalent metals, assuming one electron per atom (see Table 5.1).

We have seen that the Hall effect has a very simple explanation. The Hall voltage is due to the magnetic force on the drifting charges.

Table 5.1 Number density of conduction electrons in metals. Theoretical value of n assumes one electron per atom

Metal	Value of n based on Hall effect ($\times 10^{22}/\text{cm}^3$)	Theoretical value of n ($\times 10^{22}/\text{cm}^3$)
Lithium	3.7	4.8
Sodium	2.5	2.6
Potassium	1.5	1.3
Copper	11	8.4
Silver	7.4	6.0

The charges drift because an electric field is established between the two opposite faces R and S. It turns out that this simple explanation is not valid for non-monovalent metals (such as iron) and for semiconductors (such as germanium).

EXAMPLE 5.14 A flat silver strip of width 1.5 cm and thickness 1.5 mm carries a current of 150A. A magnetic field of 2.0T is applied perpendicular to the flat face of the strip. The Hall voltage developed across the width of the strip is measured to be 17.9 μV. Estimate the number density of free electrons in silver.

Solution: $I = 150\text{A}$

$B = 2.0\text{T}$

Width $(d) = 1.5 \times 10^{-2}$ m

Thickness $(h) = 1.5 \times 10^{-3}$ m

Hall voltage $V_\text{H} = 17.9 \times 10^{-6}$ V

$e = 1.6 \times 10^{-19}$ C

From Eq. (5.61), the number density of free electrons is given by

$$n = \frac{IBd}{V_\text{H} eA}$$

where $A = d \times h$. Substituting the numerical values of I, B, d, V_H, e and A and solving we get

$$n = 7 \times 10^{28} \text{ per m}^3$$

SUMMARY

A moving charge (i.e. current) is a source of magnetic field. The magnitude and direction of the field due to a current element is given by Biot-Savart law.

When a charge moves in a magnetic field, it experiences a magnetic force which is perpendicular to the direction of motion and the magnetic field. When a charge (stationary or in motion) is in an electric field, it experiences an electric force which is in the direction of the electric field. The two forces together constitute the Lorentz force.

Ampere's circuital law states that the line integral of magnetic field along a closed current-carrying circuit is equal to μ_0 times the current threading the circuit, irrespective of the shape of the path of integration.

A current-carrying coil suspended in a magnetic field experiences a torque. This fact is used in the construction of a galvanometer.

When a moving charge finds itself in a uniform magnetic field, it is thrown into a circular motion in a plane perpendicular to the field. A cyclotron works on this principle.

The production of a transverse voltage when a current-carrying metal strip is placed in a magnetic field is called the Hall effect. This effect confirms that the charge carriers in metals are the free electrons.

EXERCISES

A. Short Answer Questions

1. Name the basic source of a magnetic field.
2. State and define the SI unit of magnetic field.
3. State the right-hand rule for finding the direction of the magnetic field due to a long straight current-carrying conductor.
4. What is the relation between a gauss and a tesla?
5. State Fleming's left-hand rule.
6. Define the standard ampere.
7. Write down the expression for Lorentz force.
8. State Ampere's circuital Law.
9. The force **F** experienced by a particle of charge moving with a velocity **v** in a magnetic field **B** is given by $\mathbf{F} = q\ (\mathbf{v}\times\mathbf{B})$. Which pairs of vectors are always at right angles to each other? Which vectors may have any angle between them?
10. Imagine that a uniform magnetic field exists in a room where you are sitting with your back to one wall. An electron beam, travelling horizontally from the back wall toward the front wall, is deflected to your right. What is the direction of the magnetic field?
11. What is your answer to Question 10 if, instead of an electron beam, there was a proton beam?
12. If an electron is not deflected in passing through a certain region of space, can we say with certainty that there is no magnetic field in that region?
13. If a moving electron is deflected sideways in passing through a certain region of space, can we say with certainty that a magnetic field exists in that region?
14. A beam of protons can be deflected sideways by an electric field or a magnetic field. By observing the direction of the deflection, how can you tell which field is present?

B. Long-Answer Questions

1. State the Biot-Savart law and apply it to calculate the magnetic field: (i) at the centre of a circular coil carrying current and (ii) on the axis of a circular coil carrying current. Draw a rough sketch of the field lines in each case.
2. Deduce an expression for the magnetic field at a point which is at a distance of R from an infinitely long straight conductor carrying a current I. Draw a rough sketch of the field lines.
3. A uniform wire of length l subtends an angle θ with the direction of a uniform magnetic field **B**. A current I is passed through it. Obtain an expression for the magnitude of the force acting on it in terms of I, B, l and θ. What is its direction?
4. A rectangular coil is placed in a uniform magnetic field and a current is passed through it. Calculate the torque acting on it when the plane of coil is (i) parallel to the field lines, and (ii) inclined at an angle to the field lines.
5. Give the construction, theory and working of a moving coil galvanometer.
6. Obtain an expression for the force between two infinitely long parallel wires carrying current. Hence define an ampere of current.
7. State Ampere's circuital law and prove it for the case of an infinitely long current carrying thin conductor.
8. Apply Ampere's law to find the magnetic field set up by a current carrying: (a) toroidal solenoid and (b) straight solenoid. Where are such solenoids used?
9. Obtain an expression for the frequency of the circulation of a charged particle in a uniform magnetic field.
10. Describe the construction and working of a cyclotron.
11. Describe JJ Thomson's method for the measurement of charge to mass ratio of an electron.
12. What is the Hall effect? How can the effect be explained on the free-electron model of a conductor.

C. Multiple-Choice Questions

Choose the correct answer/answers from the given alternatives.

1. The force experienced by a charged particle in a magnetic field is the maximum when the particle is:

   (a) at rest
   (b) moving in a direction parallel to the field
   (c) moving in a direction perpendicular to the field
   (d) moving in a direction at an angle of 45° to the direction of the field

2. Choose the correct statement. There will be no force experienced if

   (a) two parallel wires carry currents in the same direction
   (b) two parallel wires carry currents in the opposite direction
   (c) a positive charge is projected between the pole pieces of a bar magnet
   (d) a positive charge is projected along the axis of a solenoid carrying current

3. A straight wire 2 m long carries a current of 5A. What is the force on the wire when it is placed perpendicular to a uniform magnetic field of 0.5 tesla?

(a) 5 N (b) 10 N
(c) 20 N (d) Zero

4. If an electron and a proton are projected at right angles to a uniform magnetic field with the same kinetic energy,

(a) the electron trajectory will be less curved than the proton trajectory
(b) the proton trajectory will be less curved than the electron trajectory
(c) both trajectories will be equally curved
(d) both particles move in straight lines.

5. If an electron and a proton are projected at right angles to a uniform magnetic field with the same linear momentum,

(a) the electron trajectory will be less curved than the proton trajectory
(b) the proton trajectory will be less curved than the electron trajectory
(c) both trajectories will be equally curved
(d) both particles will move in straight lines.

6. An electron is accelerated to a high speed down the axis of a cathode ray tube by the application of a potential difference of V volts between the cathode and the anode. The particle then passes through a uniform transverse magnetic field in which it experiences a force F, If the potential difference between the anode and the cathode is increased to $2\ V$, the electron will now experience a force

(a) $\frac{F}{\sqrt{2}}$ (b) $\frac{F}{2}$
(c) $\sqrt{2}\,F$ (d) $2F$

7. Which of the following in motion can NOT be deflected by magnetic fields?

(a) protons (b) cathode rays
(c) alpha particles (d) neutrons

8. A length of wire carries a steady current. It is bent first to form a circular coil of one turn. The same length is now bent more sharply to give a double loop of a smaller radius. The magnetic field at the centre caused by the same current is:

(a) a quarter of its first value (b) a half of its first value
(c) four times its first value (d) unaltered

9. Two identical coils carry equal currents and have a common centre, but their planes are at right angles to each other. What is the magnitude of the resultant magnetic field at the centre if the field due to one coil alone is B?

(a) Zero (b) $B/\sqrt{2}$
(c) $\sqrt{2}\,B$ (d) $2\ B$

10. Two long parallel wires separated by a distance R have equal current I flowing in each. The magnetic field of one exerts a force F on the other. The distance R is increased to $2R$ and the current in each wire is reduced from I to $I/2$. What is the force between them now?

(a) $4F$ (b) F
(c) $F/4$ (d) $F/8$

11. Choose the WRONG statement. The sensitivity of a moving coil galvanometer can be increased by
    (a) increasing the number of turns in the coil
    (b) inserting a soft iron cylinder inside the coil
    (c) increasing the strength of the magnetic field
    (d) using a suspension fibre of a higher restoring couple per unit twist.
12. When a charged particle moves perpendicular to a uniform magnetic field, its
    (a) energy and momentum both change
    (b) energy changes but momentum remains unchanged
    (c) momentum changes but energy remains unchanged
    (d) energy and momentum both do not change.

D. Numerical Problems

(Take $\mu_0 = 4\pi \times 10^{-7}$ TmA^{-1} wherever necessary)

1. A circular coil of 200 turns has a radius of 5 cm and carries a current of 5 A. Calculate the magnetic field: (a) at the centre of the coil, and (b) at a point at a distance of 10 cm from the centre on the axis of the coil.
2. A horizontal uniform wire of length 20 cm and mass 5.0×10^{-3} kg per meter carries a current of 2A. Find the magnitude and the direction of the magnetic field which can support the weight of the wire in air.
3. The wire shown in Fig. 5.29 carries a current of 60A. Find the magnitude and direction of magnetic field at point P.
4. A circular wire loop of radius 100 cm is carrying a current. At what distance from the centre of the loop (on its axis) is the value of magnetic field 1/10 of its value at the centre?

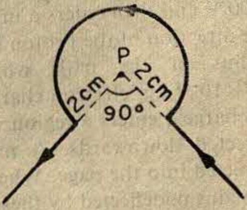


Fig. 5.29

5. A horizontal straight wire 5.0 cm long weighing 1.2 g m^{-1} is placed perpendicular to a uniform horizontal magnetic field of 0.6 T. If the resistance of the wire is 3.8 Ω m^{-1}, find the potential difference that has to be applied between the ends of the wire to make is self-supporting.
6. Two very long parallel wires carry currents of 2A and 3A in opposite directions. If they are separated by a distance of 10 cm in air, find the force per unit length between them. Is the force attractive or repulsive?
7. A long straight wire carries a current of 5.0A. An electron travels at a speed of 6.0×10^4 ms^{-1} parallel to the wire 0.10 m from it and in a direction opposite to the current. Calculate the force experienced by the electron in the magnetic field of the current.
8. Three long, straight and parallel wires, carrying currents, are arranged as shown in Fig. 5.30. Find the force experienced by a 25 cm length of wire C.
9. A 2.0 m long wire carrying a current of 5.0 A is placed in a magnetic field of 0.30 T. Find the force on the wire if it is placed: (a) at right angles to the field, (b) at 45° to the field and (c) along the field.

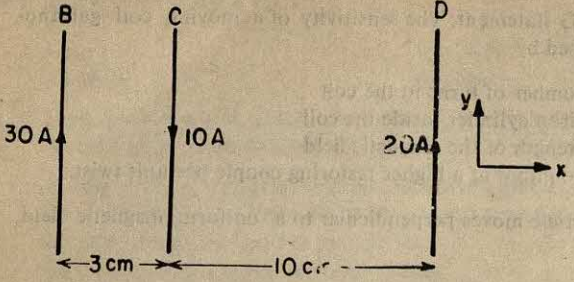


Fig. 5.30

10. Two long parallel wires carry currents of 12A and 8 A in the same direction. If the wires are 10 cm apart in air, find where a third wire, also carrying a current, must be placed so that it experiences no net force.
11. A copper wire has 1.0×10^{29} free electrons per m^3, a cross-sectional area of 6.0 mm^2 and carries a current of 5.0A. Calculate the force acting on each electron, if the wire is placed in a magnetic field of 0.15 T perpendicular to the wire.
12. A long straight wire carries a current of 50A. An electron, travelling at 1.0×10^7 ms^{-1} is 5.0 cm from the wire. Find the force acting on the electron if its velocity is directed: (a) toward the wire, (b) parallel to the wire and (c) perpendicular to the directions defined in (a) and (b).
13. A beam of protons is accelerated from rest by a potential difference of 2000 V and then enters a uniform magnetic field which is perpendicular to the direction of the proton beam. If the magnetic field is 0.2 T, calculate the radius of the path which the beam describes. Mass of a proton = 1.7×10^{-27} kg, electron charge = 1.6×10^{-19} C.
14. A charge q enters a region where a uniform electric field $E = 10^5$ V m^{-1} is directed downwards. A magnetic field of 0.4 T and perpendicular to E is directed into the page. What should be the speed of the particle so that it remains undeflected by these crossed electric and magnetic fields?
15. A charge of 1.6×10^{-19} C is moving at 1 km s^{-1} through a magnetic field in such a direction that the magnetic force on it is maximum. What is the magnetic field, if the maximum force on the charge is 8×10^{-8} N?
16. A straight conductor of length 20 cm and mass 2 g is kept horizontal and a uniform magnetic field of 10^{-2} T is applied in a horizontal direction and perpendicular to its length. Find the current to be passed through it so that its weight can be balanced by the field.
17. A rectangular coil having 1000 turns, each turn enclosing an area of 3×10^{-2} m^2 carries a current of 200 mA. It is placed such that its plane is inclined at 60° to a magnetic field of 10^{-3} T. Calculate the torque acting on the coil.
18. Considering the magnetic field along the axis of a circular loop of radius r, at what distance from the centre of the loop is the field 1/10 of its value at the centre?
19. A rectangular loop is pivoted about the y-axis and carries a current 10A in the direction shown in Fig. 5.31.

    (a) If the loop is in a uniform magnetic field of 0.2 T, parallel to the x-axis, find the force on each side of the loop and the torque required to hold the loop in the position shown.

(b) Determine the force and the torque, as in case (a), when the field is parallel to the z-axis.

(c) Find the torque required if the loop were pivoted about an axis through its centre, parallel to the y-axis.

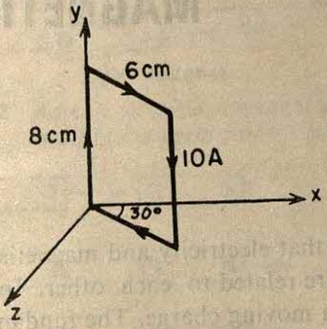


Fig. 5.31

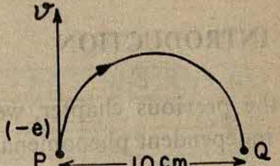


Fig. 5.32

20. An electron at point P in Fig. 5.32 has a speed of $v = 10^7 \text{ m s}^{-1}$.

(a) Find the magnitude and direction of the magnetic field that will cause the electron to follow the semi-circular path from P to Q.

(b) Find the time required for the electron to move from P to Q.

6

MAGNETISM

6.1 INTRODUCTION

In the previous chapter, we learnt that electricity and magnetism are not independent phenomena; they are related to each other. In fact, the source of a magnetic field is a moving charge. The fundamental discoveries that showed magnetism to be basically an effect of electric charges in motion were made in the early part of the 19th century by Oersted, Biot, Savart and Ampere. Historically, the knowledge of magnetism goes back to ancient times, much before these discoveries were made. The existence of natural magnetism was known to the Greeks as early as 600 BC. They found that a certain kind of rock had the ability to attract pieces of iron. This special kind of rock was called the *lodestone*. We now call it a *magnet*. The basic properties of magnets were also known and understood. They are:

1. A magnet attracts pieces of iron.
2. Every magnet has two *poles*. When a magnet is brought near a heap of iron filings, the ends of the magnet show the greatest attraction (Fig. 6.1). These ends where the magnetic attraction is the greatest, are *poles* of the magnet.

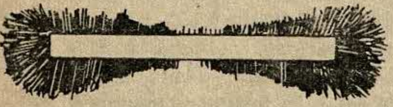


Fig. 6.1 *Poles of a magnet*

3. When a magnet is suspended or pivoted, it aligns itself in the north-south direction. The pole which points towards the geographical north is called the *north-seeking* or simply the *north pole* (N). The other pole that points towards the geographical south is called the *south-seeking* or *south pole* (S) as shown in Fig. 6.2.
4. Like poles repel and unlike poles attract. If the north poles of two magnets are brought closer, the two magnets repel each other. The force between two south poles is also repulsive. But if the north pole of one magnet is brought closer to the south pole of another magnet, they attract each other (Fig. 6.3).

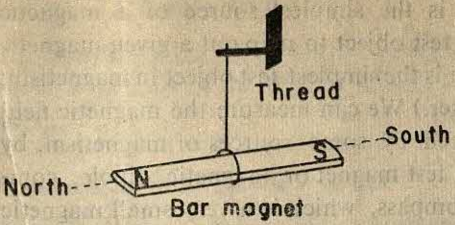


Fig. 6.2 *A freely suspended magnet aligns itself in the north-south direction*

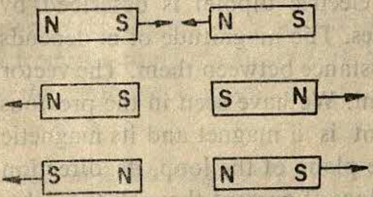


Fig. 6.3 *Like poles repel, unlike poles attract*

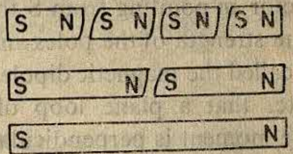


Fig. 6.4 *Poles always exist in pairs*

5. Poles always exist in pairs. If an attempt is made to separate the two poles of a magnet by breaking it in the middle, one finds new poles formed at the broken ends. If these pieces are broken again, each piece again contains new poles, a north and a south pole (Fig. 6.4). Unlike charge, a magnetic pole of one kind cannot exist without a pole of the opposite kind. Thus every magnet is a dipole.

6.2 MAGNETIC DIPOLE

In electricity the *isolated charge* is the simplest structure that can exist. If two equal charges of opposite sign are placed near each other, they form an *electric dipole*, characterized by an electric dipole moment $\mathbf{p}$. In magnetism an isolated magnetic pole which would correspond to an isolated charge does not exist. The simplest magnetic structure is the *magnetic dipole*, characterized by a magnetic dipole moment $\mathbf{m}$ as we saw in the previous chapter.

This fundamental difference between electricity and magnetism has many important consequences. For electricity, a single point charge is the simplest source of an electric field (Chapter 1). A point charge also plays the role of a *test object*. By placing the test charge at different points in an electric field and measuring the force it experiences at each point, the electric field vector can be determined at every point. Thus a given electric field is 'mapped out' by the test charge. For magnetism, a magnetic dipole is the simplest source of a magnetic field; it is also the simplest test object to map out a given magnetic field. (In fact, a moving charge is the simplest test object in magnetism; we will come to it a little later.) We can measure the magnetic field in a region of space, produced by some sources of magnetism, by measuring its effect on a small test magnet or magnetic dipole, commonly called the magnetic compass, which is a very small magnetic needle pivoted at the centre and enclosed in a brass case with a glass top (brass and glass are non-magnetic substances). A small test magnet, like its counterpart in electricity (the electric dipole) is described by a vector **m** pointing from S to N poles. The magnitude of **m** depends on the strength of the poles and the distance between them. The vector **m** is called the magnetic dipole moment. We have seen in the previous chapter that a plane loop of current is a magnet and its magnetic dipole moment is perpendicular to the plane of the loop, its direction is determined by the direction of the flow of current through the right-hand rule, and its magnitude is determined by the strength of the current and the area of the loop [Eq. (5.34) of Chapter 5].

The magnetic dipole can be used as a test object to map out a given magnetic field by measuring the torque (not force) it experiences at different points in the field (see Fig. 6.5). The torque is given by [see Eq. (5.36) of Chapter 5]

$$\boldsymbol{\tau} = \mathbf{m} \times \mathbf{B} \tag{6.1}$$

This is exactly like the torque exerted by an electric field **E** on an electric dipole of moment **p** [see Eq. (1.22)]. Under the influence of the torque, the magnet tends to align itself in the direction of the magnetic field **B** (Fig. 6.5). The reason why a freely suspended magnet aligns itself in the north-south direction is that the earth has a magnetic field of its own which exerts a torque on the suspended magnet tending to align it along its field.

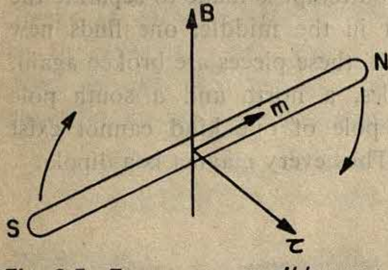


Fig. 6.5 *Torque on a small bar magnet in a magnetic field*

To summarize, a small magnet is characterized by a *vector* **m** (unlike a *scalar* charge q in electricity). The magnet experiences a *torque* in an external magnetic field (unlike a charge which experiences a *force* in an external electric field). The effect of the torque is to make the magnet align itself parallel to the external magnetic field.

SI Unit of Magnetic Moment

It follows from Eq. (6.1) that the unit of **m** is the unit of τ divided by the unit of **B**. Thus the SI unit of **m** is N m T^{-1} or J T^{-1} (joule per tesla).

Measurement of Magnetic Moment

The magnetic moment of a bar magnet can be measured by suspending it from its centre of mass and allowing it to oscillate in an external magnetic field **B**. The static equilibrium is the position when the magnet is aligned with **B**. The magnet is rotated through small angle θ from the equilibrium position and released. The magnitude of the *restoring torque* is given by

$$\tau = -mB \sin \theta$$

where m is the magnitude of the magnetic moment of the magnet. The minus sign indicates that τ is the restoring torque tending to decrease θ. For small rotation θ, $\sin \theta \simeq \theta$, so that

$$\tau = -(mB)\theta = -K\theta$$

where θ is expressed in radians and $K = mB$ is a constant. Since τ is proportional to $(-\theta)$, the condition of simple harmonic motion is met. The frequency of the simple harmonic motion is given by

$$\nu = \frac{1}{2\pi}\sqrt{\frac{K}{I}} = \frac{1}{2\pi}\sqrt{\frac{mB}{I}} \tag{6.2}$$

where I is the moment of inertia of the bar magnet about the axis of rotation. From Eq. (6.2) the value of m is given by

$$m = \frac{4\pi^2\nu^2 I}{B} \tag{6.3}$$

Thus, knowing ν, I and B, we can determine the value of m.

EXAMPLE 6.1 A bar magnet is suspended from its centre of mass and is free to rotate (about the suspension) in a plane containing a uniform magnetic field to 100 G (gauss). The magnet is displaced slightly from its stable equilibrium and released. The period of the small amplitude oscillations is measured to be 2.25s. If the moment of inertia of the magnet about its axis of rotation is 1.25×10^{-5} kg m^2, find the magnetic moment of the magnet.

Solution

$$B = 100 \text{ G} = 1.00 \times 10^{-2} \text{ T} \qquad (\because\ 1 \text{ G} = 10^{-4} \text{ T})$$

$$T = 2.25 \text{ s}$$

$$I = 1.25 \times 10^{-5} \text{ kg m}^2$$

From Eq. (6.2) the period of oscillations is given by

$$T = \frac{1}{\nu} = 2\pi \sqrt{\frac{I}{mB}}$$

$$\therefore \qquad m = \frac{4\pi^2 I}{BT^2}$$

Substituting the values of I, B and T and solving, we get

$$m = 9.75 \times 10^{-3}\ \mathrm{J\ T^{-1}}$$

EXAMPLE 6.2 A closely wound solenoid of 2000 turns and area of cross-section $1.6 \times 10^{-4}\ \mathrm{m^2}$, carrying a current of 4.0 A, is suspended through its centre allowing it to turn in a horizontal plane.

(a) What is the magnetic moment associated with the solenoid?

(b) What are the force and torque on the solenoid if a uniform horizontal magnetic field of 7.5×10^{-2} T is set up at an angle of 30° with the axis of the solenoid?

Solution: (a) The magnitude of the magnetic moment of a solenoid is given by [see Eq. (5.34) of Chapter 5 which gives **m** for one turn]

$$\begin{aligned} m &= INA \\ &= 4.0 \times 2000 \times 1.6 \times 10^{-4} \\ &= 1.28\ \mathrm{A\ m^2} \end{aligned}$$

Notice that A m² is the same as J T⁻¹. The direction of **m** is along the axis of the solenoid which is related to the sense of the current by the right-handed screw rule.

(b) The net force experienced by a magnetic dipole in a *uniform* magnetic field is zero. The torque on the solenoid is given by

$$\boldsymbol{\tau} = \mathbf{m} \times \mathbf{B}$$

The magnitude of $\boldsymbol{\tau}$ is

$$\begin{aligned} \tau &= mB \sin\theta \\ &= 1.28 \times 7.5 \times 10^{-2} \times \sin 30^\circ \\ &= 4.8 \times 10^{-2}\ \mathrm{N\ m} \end{aligned}$$

This torque tends to align the axis of the solenoid (i.e. its magnetic moment vector) parallel to field **B**.

6.3 MAGNETIC FIELD LINES

The field lines are a graphical representation of a field. The magnetic field of a bar magnet, for example, can be mapped by means of a test object (a magnetic dipole) such as a magnetic compass.

Figure 6.6(a) shows how the field lines around a bar magnet can be drawn using a compass. The bar magnet NS is placed on a sheet of paper and a compass needle is placed close to one of its poles, say the south pole S. The positions of the ends n and s of the needle are marked by pencil dots. Let the two dots be A and B. The compass is shifted so that its north pole points towards B, and the new position C of its south pole is marked. This process is repeated to get a series of dots which are then joined by a smooth curve. In the same way, other lines are drawn and each line is labelled with an arrow to indicate the direction which is from N to S.

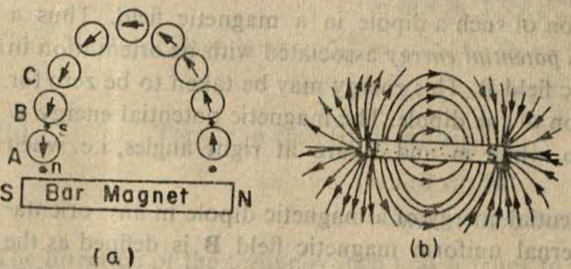


Fig. 6.6 *(a) Drawing the field lines around a bar magnet*
(b) Field lines pattern around a bar magnet

Figure 6.6(b) shows the field lines around a bar magnet. It may be mentioned that the field lines are shown only in one plane, the plane of the paper; they exist in all the space around the magnet. Compare these field lines with those of the electric field around an electric dipole [Fig. 1.23(a) of Chapter 1], the similarity of the two is evident.

Figure 6.7(a) shows the field lines around a current-carrying straight solenoid. A comparison with Fig. 6.6(b) clearly shows that such a

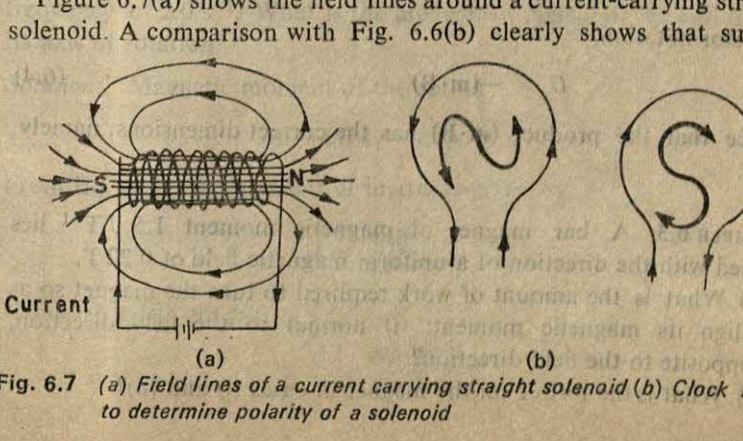


Fig. 6.7 *(a) Field lines of a current carrying straight solenoid (b) Clock rule to determine polarity of a solenoid*

solenoid behaves like a bar magnet. The polarity of any face of the solenoid can be determined following the so-called clock rule. Looking at the face, if the current around it is in the anticlockwise sense, that face has north polarity but if the current is in the clockwise sense, then it has south polarity [Fig. 6.7(b)].

6.4 POTENTIAL ENERGY OF A MAGNETIC DIPOLE IN AN EXTERNAL MAGNETIC FIELD

We have seen that when a magnetic dipole (a bar magnet or a current loop) is placed in an external magnetic field, the field exerts a torque on it. It follows that work must be done by an external agent to change the orientation of such a dipole in a magnetic field. Thus a magnetic dipole has *potential energy* associated with its orientation in an external magnetic field **B**. This energy may be taken to be zero for any arbitrary position of the dipole. The magnetic potential energy U is taken to be zero when **m** and **B** are at right angles, i.e. when $\theta = 90°$.

The magnetic potential energy of a magnetic dipole in any orientation θ with an external uniform magnetic field **B** is defined as the work that an external agent must do to turn the dipole from its zero-energy position ($\theta = 90°$) to the given position θ. Thus

$$U = \int_{\theta=90°}^{\theta} \tau \, d\theta = \int_{90°}^{\theta} mB \sin\theta \, d\theta = mB \int_{90°}^{\theta} \sin\theta \, d\theta$$

$$= -mB\cos\theta \Big|_{90°}^{\theta}$$

or
$$U = -mB\cos\theta \tag{6.3}$$

In vector notation,

$$U = -(\mathbf{m}\cdot\mathbf{B}) \tag{6.4}$$

Notice that the product $(\mathbf{m}\cdot\mathbf{B})$ has the correct dimensions, namely, joule.

EXAMPLE 6.3 A bar magnet of magnetic moment 1.5 $\mathrm{J\,T^{-1}}$ lies aligned with the direction of a uniform magnetic field of 0.22 T.

(a) What is the amount of work required to turn the magnet so as to align its magnetic moment: (i) normal to the field direction, (ii) opposite to the field direction?

(b) What is the torque on the magnet in cases (i) and (ii)?

Solution: (a) $U = -mB\cos\theta$

Potential energy when $\theta = 0$ (i.e. **m** aligned with **B**),

$$U_0 = -mB\cos 0^\circ = -mB$$

Potential energy when $\theta = 90^\circ$ (i.e. **m** perpendicular to **B**)

$$U_{90^\circ} = -mB\cos 90^\circ = 0$$

Potential energy when $\theta = 180^\circ$ (i.e. **m** opposite to **B**)

$$U_{180^\circ} = -mB\cos 180^\circ = mB$$

(i) Work done in case (i) $= U_{90^\circ} - U_0 = 0 - (-mB)$

$= mB = 1.5 \times 0.22 = 0.33$ J

(ii) Work done in case (ii) $= U_{180^\circ} - U_0 = mB - (-mB)$

$= 2mB = 0.66$ J

(b) Torque in case (i) $= mB\sin\theta$

$= mB\sin 90^\circ$

$= mB = 0.33$ Nm

The direction of the torque is such that it tends to align the magnetic moment vector **m** along the field vector **B**. Torque in case (ii) $= mB\sin 180^\circ = 0$.

EXAMPLE 6.4 A circular coil of 16 turns and radius 10 cm, carrying a current of 0.75 A, rests with its plane normal to an external field of magnitude 5.0×10^{-2} T. The coil is free to turn about an axis in its plane perpendicular to the field direction. When the coil is turned slightly and released, it oscillates about its stable equilibrium with a frequency of 2.0 s^{-1}? What is the moment of inertia of the coil about its axis of rotation?

Solution: Magnetic moment of the coil is

$$m = NiA = Ni(\pi r^2) \qquad \text{(i)}$$

From Eq. (6.3), the moment of inertia is given by

$$I = \frac{mB}{4\pi^2\nu^2} \qquad \text{(ii)}$$

Using (i) in (ii), we get

$$I = \frac{Nir^2B}{4\pi\nu^2} \qquad \text{(iii)}$$

Now $N = 16$, $r = 10$ cm $= 0.10$ m, $i = 0.75$ A, $B = 5.0\times10^{-2}$ T and $\nu = 2.0\ \text{s}^{-1}$. Putting these values in (iii) and solving, we get

$$I = 1.2\times10^{-4}\ \text{kg m}^2$$

EXAMPLE 6.5 A magnetic dipole is under the influence of two magnetic fields. The angle between the field directions is 60°, and one of the fields has a magnitude of 1.2×10^{-2} T. If the dipole comes to stable equilibrium at an angle of 15° with this field, what is the magnitude of the other field?

Solution: Let θ_1 ($= 15°$) be the angle between the magnetic moment vector **m** and the field vector $\mathbf{B}_1$ ($= 1.2\times10^{-2}$ T). Then the angle between **m** and the other field $\mathbf{B}_2$ will be $\theta_2 = 60° - 15° = 45°$.

The field $\mathbf{B}_1$ exerts a torque $\boldsymbol{\tau}_1 = \mathbf{m}\times\mathbf{B}_1$, and the field $\mathbf{B}_2$ a torque $\boldsymbol{\tau}_2 = \mathbf{m}\times\mathbf{B}_2$. The net torque is $\boldsymbol{\tau} = \boldsymbol{\tau}_1 + \boldsymbol{\tau}_2$. Now, for stable equilibrium, the net torque must be zero, which gives

$$(\mathbf{m}\times\mathbf{B}_1) + (\mathbf{m}\times\mathbf{B}_2) = 0$$

In terms of magnitudes (ignoring the signs) we have

$$B_2 = \frac{B_1 \sin\theta_1}{\sin\theta_2}$$

$$= \frac{1.2\times10^{-2}\times\sin 15°}{\sin 45°} = 4.4\times10^{-3}\ \text{T}$$

6.5 MAGNETIC ANALOGUE OF GAUSS'S THEOREM IN ELECTROSTATICS: GAUSS'S THEOREM IN MAGNETISM

Gauss's theorem in electrostatics states that the surface integral of the electrostatic field **E** over a closed surface S is equal to the total charge q inside the surface divided by the permittivity of free space, i.e. [see Eq. (2.40)]

$$\oint_S \mathbf{E}\cdot d\mathbf{S} = \frac{q}{\epsilon_0}$$

Consider now a surface enclosing an electric dipole which consists of two equal and opposite charges. These charges add up to zero so that the total charge enclosed by the surface is zero. Hence the surface integral of the electric field of a dipole over a closed surface enclosing the dipole is zero, i.e.

$$\oint_S \mathbf{E}_{\text{dipole}}\cdot d\mathbf{S} = 0 \qquad (6.5)$$

Now, unlike electric field which is produced by an isolated charge, a magnetic field is produced only by a magnetic dipole because an isolated magnetic pole does not exist. Hence the magnetic analogue of Eq. (6.5) must be as follows:

$$\oint_S \mathbf{B}\cdot d\mathbf{S} = 0 \tag{6.6}$$

This is Gauss's theorem in magnetism which states that the surface integral of magnetic field over a closed surface is zero. Equation (6.6) is one of the basic equations in electromagnetism; it is a formal way of stating the fact that isolated magnetic poles do not exist or that magnetic poles always exist in pairs or that in magnetism there is no counterpart of a free charge in electricity.

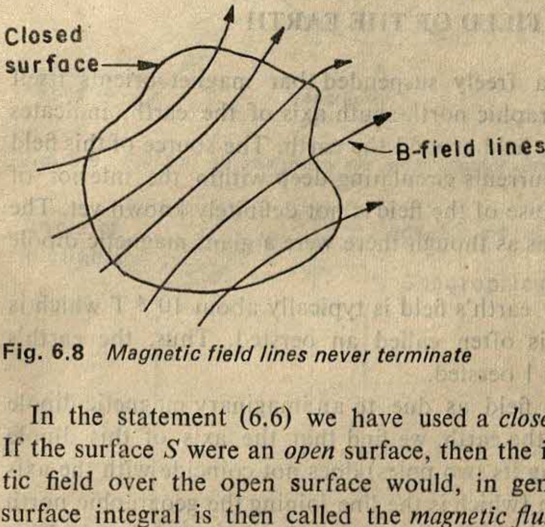


Fig. 6.8 *Magnetic field lines never terminate*

In the statement (6.6) we have used a *closed* surface S (Fig. 6.8). If the surface S were an *open* surface, then the integral of the magnetic field over the open surface would, in general, be non-zero; the surface integral is then called the *magnetic flux* (Φ) of **B** through S, i.e.

$$\Phi = \int_S \mathbf{B}\cdot d\mathbf{S} \tag{6.7}$$

The flux of a magnetic field will play an important role in Chapter 7. Because of its importance, a special unit in the SI system, named weber (symbol Wb) has been given to the flux. We, therefore, have in the SI system, in addition to the mechanical units, three electromagnetic units. They are the ampere (for current), tesla (for magnetic field strength) and weber (for magnetic flux). These units are related as follows:

$$1\ \mathrm{A} = 1\ \mathrm{J\ T^{-1}\ m^{-2}} = 1\ \mathrm{J\ Wb^{-1}}$$

$$1\ \text{T} = 1\ \text{J A}^{-1}\ \text{m}^{-2} = 1\ \text{Wb m}^{-2} \tag{6.8}$$

$$1\ \text{Wb} = 1\ \text{J A}^{-1} \quad = 1\ \text{T m}^2$$

Since tesla is the same as weber per square metre, the magnetic field strength **B** is often called the magnetic flux density and expressed in Wb m^{-2}. Thus the different alternative SI units of magnetic field strength **B**, magnetic moment **m** and magnetic permeability μ_0 are as follows:

$$\begin{aligned}(\mathbf{B}) &= \text{N A}^{-1}\ \text{m}^{-1} = \text{T} = \text{Wb m}^{-2}\\ (\mathbf{m}) &= \text{A m}^2 = \text{J T}^{-1} = \text{J m}^2\ \text{Wb}^{-1} \qquad (6.9)\\ (\mu_0) &= \text{N A}^{-2} = \text{T}^2\ \text{m}^2\ \text{N}^{-1} = \text{Wb}^2\ \text{J}^{-1}\ \text{m}^{-1}\end{aligned}$$

6.6 THE MAGNETIC FIELD OF THE EARTH

The observation that a freely suspended bar magnet orients itself roughly along the geographic north-south axis of the earth, indicates that there is a magnetic field around the earth. The source of this field is presumably electric currents circulating deep within the interior of the earth. The exact cause of the field is not definitely known yet. The earth, therefore, behaves as though there were a giant magnetic dipole embedded in it.

The strength of the earth's field is typically about 10^{-4} T which is 1 G(gauss). A gauss is often called an oersted. Thus, the earth's magnetic field is about 1 oersted.

If we visualize this field as due to an imaginary magnetic dipole embedded deep inside the earth, we find that the axis of this dipole (which is a line joining its two poles) does not coincide with the axis of rotation of the earth (which is the line joining the geographic north and south poles) but is inclined at an angle of about 20° to it. This angle changes during a time scale which is of the order of tens of thousands of years (Fig. 6.9).

Detailed charts showing the magnitude and direction of the earth's magnetic field have been prepared. They have to be revised every 30 years or so. The field of the earth is described in terms of the following three parameters.

1. Magnetic Declination

The angle between the magnetic meridian at a place and the geographic meridian at that place is called magnetic declination (Fig. 6.10). The magnetic meridian at a place is a vertical plane containing the

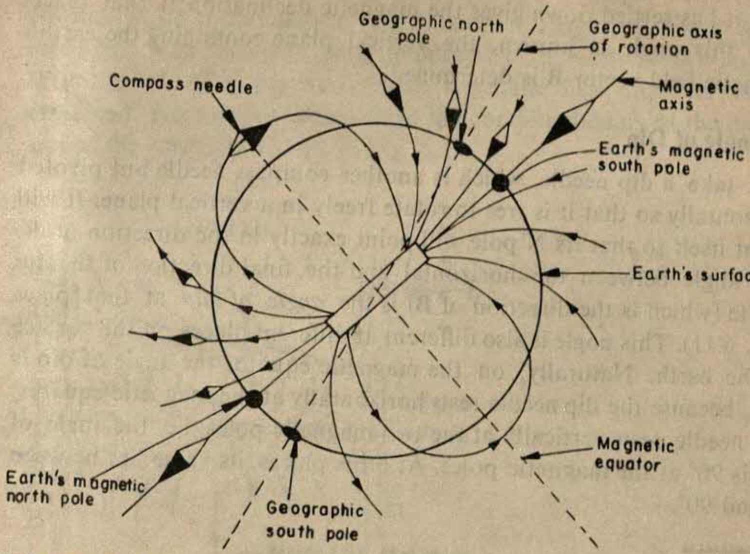


Fig. 6.9 *Field of an imaginary magnetic dipole embedded in earth*

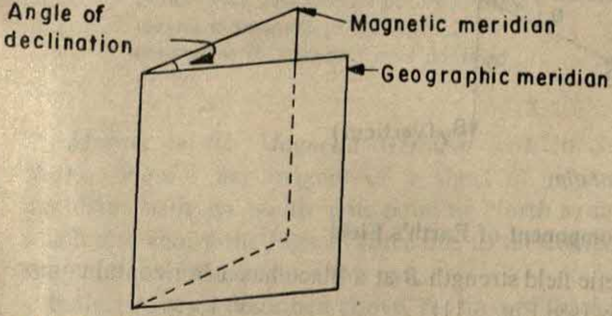


Fig. 6.10 *Magnetic declination*

magnetic axis of a freely suspended small magnet when it has settled in the earth's field. The geographic meridian at a place is a plane containing the place and the earth's axis of rotation.

The magnetic declination is different at different places on the surface of the earth. To find the direction of the magnetic field at any place, we use a *compass needle* and a *dip needle*. A compass needle is free to rotate in the horizontal plane. So, at any place in the northern hemisphere, its own N pole will finally settle down in the direction of the magnetic S pole of the earth. The angle between the true geographical north and the direction of the north pole of the compass when it has settled down gives the magnetic declination at that place. Once this angle is known, the vertical plane containing the earth's magnetic field vector **B** is determined.

2. Angle of Dip

Next take a dip needle, which is another compass needle but pivoted horizontally so that it is free to rotate freely in a vertical plane. It will orient itself so that its N pole will point exactly in the direction of **B**. The angle between the horizontal and the final direction of the dip needle (which is the direction of **B**) is the *angle of dip* at that place (Fig. 6.11). This angle is also different at different places on the surface of the earth. Naturally, on the magnetic equator the angle of dip is zero, because the dip needle rests horizontally at the magnetic equator. The needle rests vertically at the two magnetic poles, i.e. the angle of dip is 90° at the magnetic poles. At other places, its value lies between 0° and 90°.

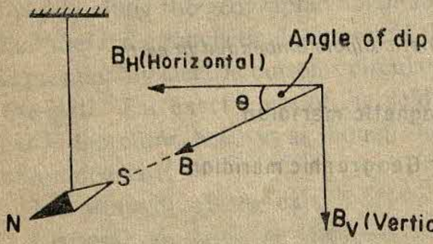


Fig. 6.11 *Angle of dip*

3. Horizontal Component of Earth's Field

The total magnetic field strength B at a place has a horizontal component B_H given by (see Fig. 6.11)

$$B_H = B \cos \theta$$

where θ is the dip angle at that place. At the magnetic poles, $B_H = B \cos 90° = 0$ and at the magnetic equator $B_H = B \cos 0° = B$. The value of B_H thus differs from place to place on the surface of the earth.

Neutral Points When a magnet is placed on a table, the magnetic field around the magnet is a combination of the magnetic field of the magnet and that of the earth. The resultant field when a magnet's field is combined with that of the earth depends on the direction in which the magnet is lying. We shall draw the lines of force in the following two special cases.

(a) *Magnet in the Magnetic Meridian with its North Pole Pointing North*: Place a bar magnet on a sheet of paper, in the magnetic meridian, with its north pole pointing North as shown in Fig. 6.12. With the help of a compass needle trace the field lines as already explained. The lines of force due to the combined fields of the magnet and of the earth are shown in Fig. 6.12.

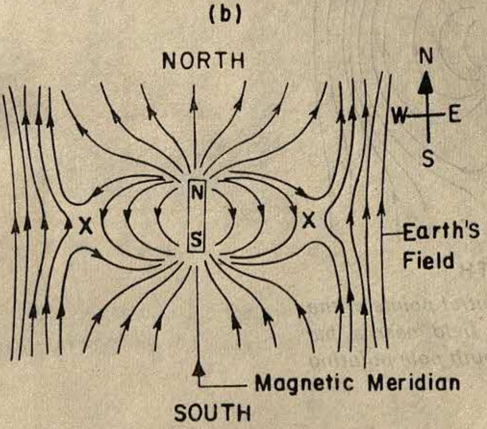


Fig. 6.12 *Field lines and neutral points of the resultant magnetic field near a bar magnet with its north pole pointing north*

(b) *Magnet in the Magnetic Meridian with its South Pole Pointing North*: Place a bar magnet on a sheet of paper, in the magnetic meridian, with its south pole pointing North as shown in Fig. 6.13, which also shows the lines of force due to the combined fields of the magnet and of the earth in this case.

In the two cases described above, the lines of force in the neighbourhood of the magnet are mainly due to the field of the magnet which is much stronger than the field of the earth. As the distance from the magnet increases, the field due to it becomes weak, until at very distant points, the field of the magnet becomes much weaker than earth's field and the lines obtained are mainly due to the earth's field. These lines are nearly straight and parallel running from South to North as shown in Figs. 6.12 and 6.13. There are two points marked X, X in the two diagrams where the field of the magnet and that of the earth are equal and opposite to each other. The resultant field at these points is, therefore, zero. These points are called neutral points. If a compass needle is placed at any of these points, no force acts on it and it may be made to set itself in any direction.

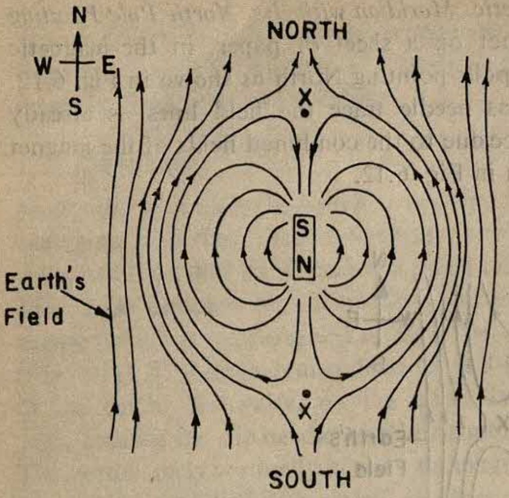


Fig. 6.13 *Field lines and neutral points of the resultant magnetic field near a bar magnet with its south pole pointing north*

Aurora Borealis

Aurora Borealis—or *Northern Lights*—is a beautiful display of colours seen in extreme northern latitudes. They are caused by the earth's magnetic field as streams of electrons rushing towards the earth are acted upon by the earth's magnetic field.

The Tangent Galvanometer

In the previous chapter we learnt how a moving coil galvanometer works. We now describe another kind of instrument, called the *tangent galvanometer* whose working is based on the existence of the earth's magnetic field. This instrument can also be used for measuring electric current.

The tangent galvanometer consists of a circular coil at the centre of which is pivoted a magnetic compass free to rotate in a horizontal plane. The coil is placed in a vertical plane containing the earth's magnetic field, i.e. the plane of the coil is made to lie in the magnetic meridian [Fig. 6.14(a)]. The current to be measured is passed through the coil.

When there is no current in the coil, the compass needle points in the north-south direction, i.e. it rests in the same vertical plane as the coil, due to the action of the horizontal component B_H of the earth's field. When a current I is passed through the coil, a magnetic field **B**

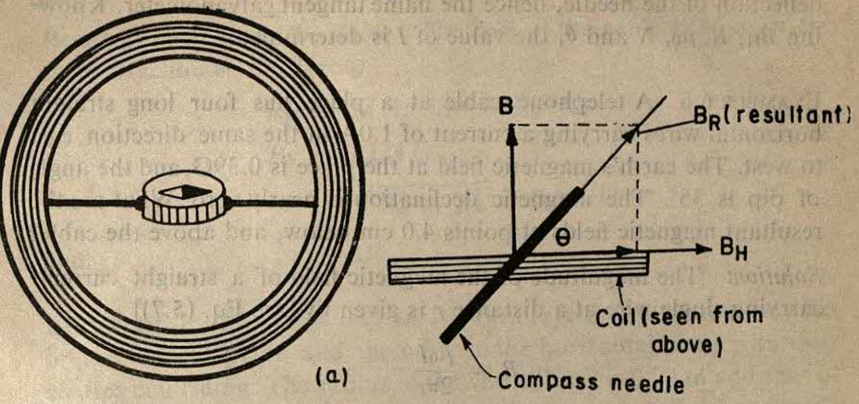


Fig. 6.14 *The tangent galvanometer* (b)

is set up around it. According to Eq. (5.10), the magnitude of the magnetic field at the centre of the coil is given by

$$B = \frac{\mu_0 NI}{R} \tag{6.10}$$

where N is the number of turns in the coil and R is its mean radius. The direction of this field is perpendicular to the plane of the coil [see Fig. 6.14(b)]. This is evidently also perpendicular to the horizontal component $\mathbf{B}_{\mathrm{H}}$ of the earth's field as shown in the diagram. The magnetic needle is acted upon by two fields; B (due to the current carrying coil) and B_{H} (of the earth). These fields are perpendicular to each other. The compass needle will rotate due to the torques of these fields and come to rest at an angle, say θ, with the plane of the coil (i.e. with the field B_{H}). It follows from Fig. 6.14(b) that B, B_{H} and θ are related as

$$\frac{B}{B_{\mathrm{H}}} = \tan\theta$$

or

$$B = B_{\mathrm{H}} \tan\theta$$

Using this expression in Eq. (6.10) we get

$$I = \left(\frac{2RB_{\mathrm{H}}}{N\mu_0}\right)\tan\theta$$

or

$$I = K\tan\theta \tag{6.11}$$

where $K = 2RB_{\mathrm{H}}/(N\mu_0)$ is a constant for the particular location of the experiment. Remember that B_{H} changes from place to place. Notice from Eq. (6.11) that the current is proportional to the tangent of the deflection of the needle, hence the name tangent galvanometer. Knowing B_H, R, μ_0, N and θ, the value of I is determined.

EXAMPLE 6.6 A telephone cable at a place has four long straight horizontal wires carrying a current of 1.0A in the same direction east to west. The earth's magnetic field at the place is 0.39G, and the angle of dip is 35°. The magnetic declination is nearly zero. What are the resultant magnetic fields at points 4.0 cm below, and above the cable?

Solution: The magnitude of the magnetic field of a straight current-carrying single wire at a distance r is given by [see Eq. (5.7)]

$$B = \frac{\mu_0 I}{2\pi r}$$

For a cable having N wires, the field is multiplied N times and is given by

$$B_C = \frac{\mu_0 N I}{2\pi r}$$

$$= \frac{(4\pi \times 10^{-7}) \times 4 \times 1.0}{2\pi \times (4.0 \times 10^{-2})}$$

$$= 0.2 \times 10^{-4}\ \mathrm{T} = 0.2\ \mathrm{G}$$

The horizontal and vertical components of the earth's field **B** are (see Fig. 6.11)

$$B_H = B\cos\theta = 0.39\cos 35° = 0.319\ \mathrm{G}$$

and $$B_V = B\sin\theta = 0.39\sin 35° = 0.224\ \mathrm{G}$$

where θ is the dip angle.

Resultant field Below the Cable The field due to the cable is in the horizontal direction but opposite to B_H. Hence the horizontal component of the resultant field is

$$R_H = B_H - B_C = 0.319 - 0.2 = 0.119\ \mathrm{G}$$

The vertical component remains unchanged and it is

$$R_V = B_V = 0.224\ \mathrm{G}$$

Therefore, the magnitude of the resultant field is

$$R = \sqrt{R_H^2 + R_V^2} = 0.254\ \mathrm{G}$$

and its direction is given by

$$\tan\alpha = \frac{R_V}{R_H} = \frac{0.224}{0.119} = 18.8$$

giving $\alpha \simeq 62^\circ$

Resultant Field above the Cable The field of the cable here is in the same direction as B_H. Hence

$$R_H = B_H + B_C = 0.519 \text{ G}$$

$$R_V = 0.224 \text{ G}$$

$$R = \sqrt{R_H^2 + R_V^2} = 0.566 \text{ G}$$

and $\alpha = 23^\circ$

EXAMPLE 6.7 Two coils each of 100 turns are held such that one lies in the vertical plane and the other in the horizontal plane with their centres coinciding. The radius of the vertical coil is 0.20 m and that of the horizontal coil is 0.30 m. How would you neutralize the magnetic field of the earth at their common centre? What is the current to be passed through each coil? Horizontal component of earth's magnetic field = 0.35×10^{-4} tesla and angle of dip = 30°.

Solution: Angle of dip $\theta = \tan^{-1}\left(\frac{B_V}{B_H}\right)$ where B_V and B_H are the vertical and horizontal components of earth's field respectively.
Since $B_H = 0.35 \times 10^{-4}$ T, we have

$$B_V = B_H \tan 30^\circ$$

$$= 0.35 \times 10^{-4} \times 0.577$$

$$= 0.202 \times 10^{-4} \text{ T}$$

The direction of the magnetic field will be perpendicular to the plane of either coil. If the plane of the vertical coil is perpendicular to the magnetic meridian, the field produced by it can neutralize the horizontal component of earth's field if

$$B_H = \frac{\mu_0}{4\pi} \cdot \frac{2\pi NI}{R} = 0.35 \times 10^{-4}$$

or

$$I = \frac{0.35 \times 10^{-4} \times 0.20 \times 10^7}{2\pi \times 100}$$

$$= 0.11 \text{ A}$$

Similarly, the magnetic field produced by the horizontal coil will be vertical and will neutralize the vertical component B_V of the earth's field, if

$$B_V = \frac{\mu_0}{4\pi} \cdot \frac{2\pi NI}{R} = 0.202 \times 10^{-4}$$

which gives $I = 0.096$ A

SUPPLEMENT

MAGNETIC PROPERTIES OF MATERIALS: DIA- PARA- AND FERROMAGNETISM

It is necessary to know the magnetic properties of different materials in order to decide which materials are suitable for loudspeakers (where permanent magnets are required) or as cores in electromagnets (where temporary magnets are required) or in transformers etc. Different materials respond differently when placed in an external magnetic field.

We begin by defining a few terms which are used to describe the magnetic properties of materials.

1. Magnetic Field in a Magnetic Material

In order to measure the magnetic effect of matter, a fixed current is passed through a toroidal solenoid. The magnetic field in the empty solenoid is then compared with the field when the solenoid is filled with a magnetic material. The difference in these two measurements gives the magnetic contribution of the specimen.

Consider a toroidal solenoid with n turns per unit length, carrying a current I wound round a ring of a magnetic material (Fig. 6.15).

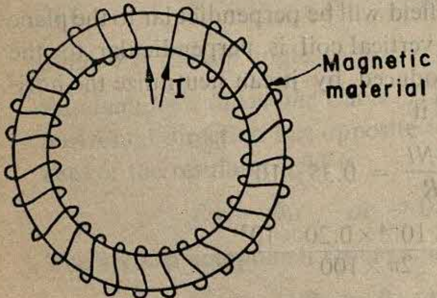


Fig. 6.15 *Magnetic field in a magnetic material*

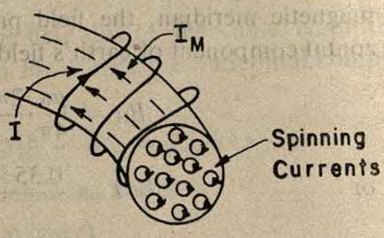


Fig. 6.16 *Magnetising surface current*

If the material is absent (i.e. we have vacuum), the magnetic field in the solenoid is given by [see Eq. (5.31)]

$$B = \mu_0 nI \tag{6.12}$$

When the experiment is repeated with a specimen filling the solenoid, the field **B** is modified by the matter. To account for the presence of matter, Eq. (6.12) is rewritten as

$$B = \mu_0 n(I + I_M)$$

or

$$B = B_0 + B_M \tag{6.13}$$

where $B_0 = \mu_0 nI$ is the field due to the current in the solenoid and $B_M = \mu_0 nI_M$ is the field due to the magnetization of the matter; I_M is the magnetizing surface current which describes the magnetic effect of the matter.

In all materials, the magnetic effects are ultimately traceable to the behaviour of electrons in the atoms of the material. When an electron is circling the nucleus of the atom, it is effectively a current loop (a circulating charge) which is equivalent to a tiny magnetic dipole (as we saw in Chapter 5) with a certain magnetic moment. The magnetic moment of an atom may be due to three factors—electron spin, the orbital angular momentum of electrons about the nucleus and the change in the orbital moment induced by an applied magnetic field. Thus the magnetic field B_M is produced by many small circulating currents inside the magnetic material due to the orbiting and spinning electrons in the atoms. Figure 6.16 shows that the effect of many adjacent current loops may be thought of as one current loop. In the same way, the internal circulating and spinning currents, due to orbiting and spinning electrons, can be replaced by a single current I_M. This surface magnetization current is in addition to the actual current I flowing in the solenoid.

2. Intensity of Magnetization

The *magnetic moment per unit volume* is called the *intensity of magnetization* or simply *magnetization M* of the material. Now the magnetic moment of each turn of the solenoid = $I_M A$, where A is the area of each turn. If n is the number of turns per unit length, the magnetic moment of the solenoid per unit length is $nI_M A$. Dividing it by area A gives the magnetic moment per unit volume.

Hence

$$M = \frac{nI_M A}{A} = nI_M$$

Now

$$B_M = \mu_0 nI_M = \mu_0 M \tag{6.14}$$

3. Magnetizing Field Intensity

The product nI is called the *magnetizing field intensity H*.

$$\therefore \qquad B_0 = \mu_0 nI = \mu_0 H \qquad (6.15)$$

Using Eqs (6.14) and (6.15) in Eq (6.13) we have

$$B = \mu_0(H + M) \qquad (6.16)$$

Equation (6.16) relates the total magnetic field B to the magnetizing field intensity H of the current and to the magnetization M produced in the material. *Note that H and M have the same units, namely, ampere per metre (Am^{-1}).*

4. Permeability

It turns out that the total magnetic field B of a material is proportional to the magnetizing field intensity H due to the current flowing in the solenoid, i.e.

$$B \propto H$$

$$\text{or} \qquad B = \mu H \qquad (6.17)$$

The constant μ is called the *permeability* of the material. If the material is removed leaving a vacuum inside the solenoid, the magnetic field reduces to B_0 given by

$$B_0 = \mu_0 H$$

where μ_0 is the permeability of vacuum.

The ratio μ/μ_0 is called the relative permeability of the material.

$$\mu_r = \frac{\mu}{\mu_0} \qquad (6.18)$$

From Eqs (6.16) and (6.17), we have

$$\mu H = \mu_0(H + M)$$

$$\text{or} \qquad \frac{\mu}{\mu_0} = 1 + \frac{M}{H}$$

$$\text{or} \qquad \mu_r = 1 + \frac{M}{H} \qquad (6.19)$$

5. Susceptibility

For small values of H, M is found to be proportional to H, i.e.

$$M \propto H$$

$$\text{or} \qquad M = \chi H \qquad (6.20)$$

where χ is a constant called the susceptibility of the material and is a dimensionless number like μ_r. Using Eq. (6.20) in Eq. (6.19) gives

$$\mu_r = 1 + \chi \qquad (6.21)$$

THREE KINDS OF MAGNETIC MATERIALS

Different magnetic materials have different χ values which may vary over a very wide range. Depending on the value and sign of χ, magnetic materials are divided into the following three categories.

1. Diamagnetic Materials

Materials for which χ is small and negative are called *diamagnetic.* In such materials, the individual magnetic moments of the various atoms tend to cancel out completely. So, in the normal state, the atoms of such substances (in the absence of external field) have no magnetic moment at all. They have a small and negative susceptibility χ. For bismuth $\chi = -0.00015$ and $\mu_r(=1 + \chi)$ is slightly less than unity. Bismuth, copper, lead, water and sodium chloride are all diamagnetic.

When a sample of a diamagnetic material is placed in a magnetic field **B**, a net magnetic moment is induced in it which is proportional to **B** but opposite in direction. If **B** is non-uniform, the sample tends to move from the region where the field is strong to the region where it is weak. It should be noted that diamagnetism is a natural reaction to an applied magnetic field and this behaviour is temperature-independent.

2. Paramagnetic Materials

Paramagnetic materials are those whose atoms, in the normal state, already have an intrinsic nonzero magnetic moment, even in the absence of an applied magnetic field. Examples are aluminium, sodium, copper chloride and liquid oxygen. Paramagnetics have a small but positive value of χ. The susceptibility χ satisfies the relation

$$\chi = \frac{C}{T} \tag{6.22}$$

where C is a constant called the *Curie constant* and the relation (6.22) is called *Curie's law*.

When a sample of a paramagnetic material is placed in a magnetic field **B**, the individual magnetic moments tend to align themselves in the direction of **B** while thermal motions tend to destroy this alignment. If **B** is strong enough and the temperature is low enough, there is a net average magnetic moment density pointing in the same direction as **B**. So such a sample will move from a low-field region to a high-field region, exactly opposite to the behaviour of diamagnetic materials. Paramagnetism is obviously temperature-dependent.

3. Ferromagnetic Materials

Materials having a very high (of the order of 1000 or more) and positive susceptibility are called ferromagnetic. Examples are iron, nickel, cobalt and numerous alloys. In such materials there are strong interactions between the individual magnetic moments of the various atoms. These interactions can be understood only in terms of the quantum theory, which is outside the scope of this book. These strong interactions cause the neighbouring moments to align even in the absence of the applied field. These alignments take place in bulk volume leading to a spontaneous bulk magnetization of these materials. This is the origin of macroscopic magnets which may be natural (such as lodestone) or artificially made permanent magnets. Above a certain temperature called the *Curie point*, ferromagnetics becomes paramagnetics.

Hysteresis: In ferromagnetic materials (e.g. soft iron) the relation $B = \mu H$ does not hold, i.e. the magnetic field B of the material does not increase linearly with the magnetizing field H. The relation between them is quite complicated; it can even depend on the past history of the sample.

Take an unmagnetized rod of soft iron and wind round it a large number of turns of wire. We have a straight solenoid with a soft iron core. Pass a current I through the solenoid. The value of I and the number of turns per unit length (n) determine the magnetizing field $H(=nI)$. Increase the value of current I gradually. It is found that the magnetic field B (which depends on magnetization M) increases gradually as H is increased. B is much larger than H because the current I (which is small initially) produces a moderate H field which has the effect of aligning the small atomic magnetic moments in its own direction, thus leading to a large M and hence a large B-field. As current I (and hence H) is, increased further, a stage is reached when M saturates, i.e. it becomes a constant and does not increase any more with increasing H. In other words, the B-field saturates at a certain value of the H-field (see Fig. 6.17).

If now the H-field is gradually decreased to zero (by gradually decreasing the current in the solenoid), it is found that some magnetization M (and hence the B-field) remains in the iron and the B-field does not fall to zero. In one complete cycle of change of H, the B-field follows a *closed loop* PQRS which is called the *hysteresis loop*. The fact that for each value of H, there is no unique value of B, but a value dependent on the manner or history by which that H was reached, is called the phenomenon of *hysteresis*. Hysteresis, which comes

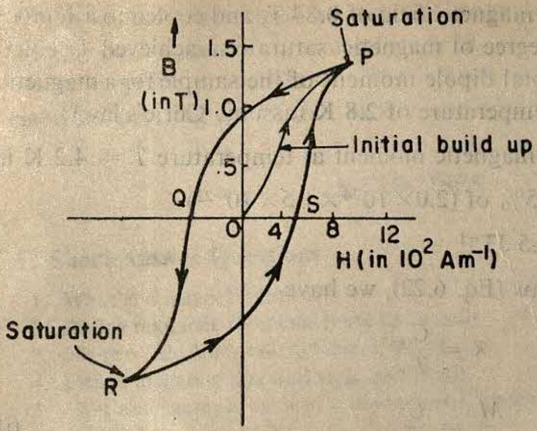


Fig. 6.17 *Hysteresis loop*

from a Greek word meaning 'delayed' can be defined as the lagging of *B* behind the magnetizing field *H*, when the specimen is taken through a complete magnetic cycle.

EXAMPLE 6.8 A toroidal solenoid has 3000 turns and a mean radius of 10 cm. It has a soft iron core of relative permeability 2000. Find the magnitude of the magnetic field in the core when a current of 1.0 A is passed through the solenoid.

Solution: The magnitude of the magnetic field in the core is given by

$$B = \mu n I$$

or

$$B = \mu_r \mu_0 n I \qquad (\because \quad \mu_r = \mu/\mu_0) \qquad \text{(i)}$$

where n is the number of turns per unit length. If r is the mean radius (given to be 10 cm = 0.10 m), the value of n is (since total length of the solenoid is $2\pi r$)

$$n = \frac{3000}{2\pi r}$$

Now $\mu_r = 2000$, $\mu_0 = 4\pi \times 10^{-7}$ TmA^{-1} and $I = 1.0$ A. Substituting these values in (i) we have

$$B = 12 \text{ T}$$

Notice that the field is increased considerably by inserting a core of a high relative permeability.

EXAMPLE 6.9 A sample of paramagnetic salt contains 2.0×10^{24} atomic dipoles each of dipole moment 1.5×10^{-23} JT^{-1}. The sample is placed under a homogeneous magnetic field of 0.84 T, and cooled to a temperature of 4.2 K. The degree of magnetic saturation achieved is equal to 15%. What is the total dipole moment of the sample for a magnetic field of 0.98 T and a temperature of 2.8 K (assume Curie's law)?

Solution: Initial total magnetic moment at temperature $T = 4.2$ K is

$$m_1 = 15\% \text{ of } (2.0 \times 10^{24} \times 1.5 \times 10^{-23})$$

$$= 4.5 \text{ JT}^{-1}$$

Now from Curie's law (Eq. 6.22), we have

$$\chi = \frac{C}{T}$$

or

$$\frac{M}{H} = \frac{C}{T} \qquad \text{(i)}$$

where M is magnetization and H is the magnetizing field. If V is the volume of the sample, then, by definition $m = M/V$ and Eq. (i) becomes

$$m = CV\left(\frac{H}{T}\right) = \text{constant} \times \left(\frac{H}{T}\right) \qquad \text{(ii)}$$

If m_2 is the magnetic moment at temperature $T_2 = 2.8$ K and field $H_2 = 0.98$ T, then from Eq. (ii) m_1 and m_2 are related as

$$\frac{m_1}{m_2} = \frac{H_1}{H_2} \cdot \frac{T_2}{T_1}$$

or

$$m_2 = m_1 \times \frac{H_2}{H_1} \times \frac{T_1}{T_2}$$

$$= 4.5 \times \frac{0.98}{0.84} \times \frac{4.2}{2.8} = 7.9 \text{ JT}^{-1}$$

SUMMARY

Magnetic dipole is the simplest structure in magnetism. It experiences a torque when placed in a magnetic field, tending it to align itself parallel to the field. It therefore has magnetic potential energy at any general orientation with the field.

Gauss's theorem in magnetism states that the surface integral of the magnetic field over a closed surface is zero.

The earth has a magnetic field of its own which is caused presumably by electric currents circulating deep within its interior. The field of the earth is described in terms of magnetic declination, dip angle and the horizontal component of its field.

The magnetic properties of materials are ultimately traceable to the electrons orbiting round the nucleus and spinning about an axis.

EXERCISES

A. Short-Answer Questions

1. What is a magnetic dipole?
2. Define magnetic moment. State its SI unit.
3. No two field lines ever intersect. Why?
4. Define magnetic flux and state its SI unit.
5. Give one example each of a diamagnetic, paramagnetic and ferromagnetic material.

B. Long-Answer Questions

1. 'A magnetic dipole experiences a torque in a magnetic field, and a point charge experiences a force in an electric field'. Explain.
2. Define magnetic moment of a magnet. How is it experimentally measured?
3. (a) State two important properties of magnetic field lines.
   (b) Draw a rough sketch of the field pattern of a current-carrying straight solenoid: (i) without and (ii) with a soft iron core. Hence explain why such a core is used.
4. Deduce the expression for the potential energy of a magnetic dipole in an external uniform magnetic field.
5. Compare and contrast the Gauss's theorems in electrostatics and in magnetism.
6. (a) What is the earth's magnetic field due to?
   (b) Define the terms magnetic declination and dip angle.
7. Describe the construction and theory of a tangent galvanometer? Why is the instrument so called?
8. 'Magnetic properties of materials are ultimately due to the behaviour of electrons in the atoms.' Comment.
9. Explain clearly what is meant by the terms magnetization and magnetizing field intensity.
10. (a) Define the terms permeability, relative permeability and susceptibility of a substance.
    (b) What is the relation between relative permeability and susceptibility of a substance.
11. Distinguish between diamagnetic, paramagnetic and ferromagnetic substances. Give two examples of each.
12. (a) What is hysteresis?
    (b) Draw a typical hysteresis loop. What conclusions can you draw from it?

C. Multiple-Choice Questions

Choose the correct answer/answers from the given alternatives.

1. When a magnetic dipole is placed in a uniform magnetic field, it will experiences

   (a) a force but no torque (b) a torque but no force
   (c) a force as well as a torque (d) neither a force nor a torque

2. In the SI system, magnetic moment may be expressed in

   (a) TJ^{-1} (b) JT^{-1}
   (c) $A\ m^{-1}$ (d) $N\ m\ T^{-1}$

3. The lines of force due to the horizontal component of the earth's magnetic field are

   (a) straight and parallel (b) concentric circles
   (c) elliptical (d) parabolic

4. The vertical component of the earth's magnetic held is zero at the

   (a) magnetic poles (b) magnetic equator
   (c) geographic poles (d) 45° latitude

5. The angle of dip is 90° at the

   (a) magnetic poles (b) magnetic equator
   (c) geographic poles (d) 90° latitude

6. A neutral point in the magnetic field, due to a magnet in a laboratory on earth, is a point where

   (a) the lines of force of the magnet intersect each other
   (b) the earth's magnetic field is zero
   (c) the field due to the magnet is zero
   (d) the fields of the magnet and the earth are equal and opposite

7. A dip needle free to move in a vertical plane perpendicular to the magnetic meridian will remain

   (a) vertical
   (b) horizontal
   (c) at an angle of 60° to the vertical
   (d) at an angle of 45° to the horizontal

8. When a material is subjected to a small magnetic field H, the intensity of magnetisation is proportional to

   (a) $H^{1/2}$ (b) H
   (c) H^2 (d) $H^{-1/2}$

9. The magnetic permeability of a paramagnetic substance is

   (a) small and positive (b) small and negative
   (c) large and positive (d) large and negative

10. Which of the following has the highest magnetic permeability?

    (a) paramagnetic substances (b) diamagnetic substances
    (c) ferromagnetic substances (d) vacuum

11. When a ferromagnetic substance is heated to a temperature above its Curie temperature, it

(a) behaves like a paramagnetic substance
(b) behaves like a diamagnetic substance
(c) remains ferromagnetic (d) is permanently magnetised

12. The relative permeability of iron is of the order of
(a) zero (b) 10^{-4}
(c) 1 (d) 10^{4}

13. The most suitable metal for making permanent magnets is
(a) iron (b) steel
(c) copper (d) aluminium

14. The most suitable metal for making electromagnets, cores of transformers and armatures in machines is
(a) iron (b) steel
(c) copper (d) silver

15. Large transformers, when used for some time, become very hot and are cooled by circulating oil. The heating of the transformers is due to
(a) the heating effect of the current alone
(b) hysteresis loss alone
(c) both the hysteresis loss and heating effect of the current
(d) none of the above

16. The relative permeability of a substance A is slightly greater than unity while that of a substance B is slightly less than unity. Then
(a) A is ferromagnetic and B paramagnetic
(b) A is diamagnetic and B paramagnetic
(c) A is paramagnetic and B diamagnetic
(d) A and B are both ferromagnetic

D. Numerical Problems

1. A small bar magnet of magnetic moment 6.7×10^{-2} JT^{-1} is suspended from its centre of gravity and is free to rotate in a plane containing a uniform magnetic field. It is displaced from its stable equilibrium and the period of small oscillations is found to be 2/3 s. If the moment of inertia of the magnet about its axis of rotation is 7.5×10^{-6} kgm^2, determine the magnitude of the magnetic field.
2. A short bar magnet of magnetic moment 0.40 JT^{-1} is placed with its axis at 30° with a uniform external magnetic field of 0.16 T. Find the magnitude of the torque exerted on the magnet by the field.
3. A straight solenoid of length 50 cm has 1000 turns and a mean cross-sectional area of 2.0×10^{-4} m^2. It is placed with its axis at 30° with a uniform magnetic field of 0.32 T. Find the torque acting on the solenoid when a current of 2.0 A is passed through it.
4. A short bar magnet of magnetic moment 5.25×10^{-2} JT^{-1} is placed with its axis perpendicular to the earth's field direction. At what distance from the centre of the magnet on:

   (i) its normal bisector,
   (ii) its axis

is the resultant field inclined at 45° with the earth's field. Magnitude of the earth's field at the place is given to be 0.42 G. Ignore the length of the magnet in comparison to the distances involved.

5. A Rowland ring of mean radius 15 cm has 3500 turns of wire wound on a ferromagnetic core of relative permeability 800. What is the magneticfield (B) in the core for a magnetizing current of 1.2 A?
6. A short bar magnet is placed in a horizontal plane with its axis in the magnetic meridian. Neutral points are found on its normal bisector at 10.0 cm from the centre of the magnet. The earth's magnetic field at the place is 0.35 G and the dip angle is zero.
   (a) Find the total magnetic field at a point on the axis of the magnet at a distance of 10.0 cm from its centre.
   (b) Locate the neutral points when the bar magnet is turned around by 180°. Neglect the length of the magnet.
7. A magnetic dipole is under the influence of two magnetic fields $B_1 = 2.2\times10^{-2}$ T and $B_2 = 4.4\times10^{-3}$ T inclined to each other at an angle of 60°. Find the position of the stable equilibrium of the dipole.
8. A soft iron ring has a mean diameter of 20 cm and a cross-sectional area of 5.0×10^{-4} m^2. It is uniformly wound with 2000 turns and carries a current of 2.0 A. If the magnetic flux in the iron is 8.0×10^{-3} Wb, what is the relative permeability of iron.
9. A circular coil of radius 0.05 m, having 40 turns, lies in a horizontal plane. It carries a current of 5.0 A in anti-clockwise sense as seen from above. The coil is placed in a magnetic field of 1.2 T directed at right angles to its axis. Find the magnitude of the torque on the coil.
10. A paramagnetic sample has a magnetic moment of 5.0 JT^{-1} at a temperature of 4.0 K and a magnetic field of 1.0 T. What will be its magnetic moment if the temperature is decreased to 3.0 K and magnetic field increased to 1.5 T?

7

ELECTROMAGNETIC INDUCTION

7.1 INTRODUCTION: PHENOMENON OF ELECTROMAGNETIC INDUCTION

After Oersted's discovery that an electric current produces a magnetic field, attempts were made to observe the reverse effect, i.e. generation of electric current by means of a magnetic field. Joseph Henry (1830) in the USA and Michael Faraday (1831) in UK independently observed that whenever there is a change in the magnetic lines of force associated with a conductor, an electromotive force (emf) is set up at the ends of the conductor which lasts as long as the change is taking place. This phenomenon of generating emf is called *electromagnetic induction*. The emf developed is called *induced emf*. If the conductor is in the form of a closed circuit, a current flows in the circuit. This current is called *induced current*. This phenomenon and the laws governing it are of great practical value in daily life. This basic phenomenon is the basis of the working of a power generator, dynamo, transformer, etc. because after its discovery it was possible to convert electrical energy into mechanical energy with high efficiency.

The phenomenon of electromagnetic induction can be demonstrated by a few simple experiments conducted by Faraday and others. Figure 7.1a shows a circular insulated wire of one or more turns connected to a sensitive galvanometer G. If a bar magnet NS is moved towards the coil as shown by the arrow, the galvanometer needle deflects, showing the presence of an electric current in the circuit. Similarly, if the magnet is moved away from the coil, the deflection is in the opposite direction. When the motion ceases, there is no deflection in the galvanometer. Further, if the magnet is moved slowly, smaller deflection results as compared to the case when it is moved faster. The motion of the magnet implies that the number of magnetic lines of force entering the coil is changing and it is constant when the motion ceases. Further, the rate of change of lines of force is less when the magnet moves slowly while it is more when it moves faster. The results of this experiment may be summarized as follows:

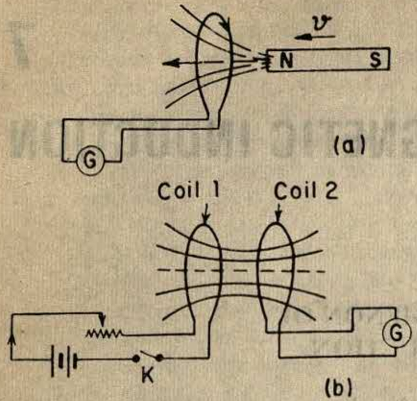


Fig. 7.1 *Demonstration experiments of electromagnetic induction*

1. A current flows in the circuit when the number of magnetic lines of force associated with a coil or loop of wire is changing.

2. The higher the rate of change of lines of force, the greater is the current.

The same effect can also be observed when the magnetic field is produced by passing a current in coil 1 and connecting the sensitive galvanometer G to coil 2 placed near coil 1, as shown in Fig. 7.1b. As soon as the key K is inserted in the circuit, there is a momentary deflection in the galvanometer; similarly, when K is removed from the circuit, there is again a momentary deflection in G. The whole process can be understood on similar lines as in the previous experiment. As soon as K is inserted, i.e. when the circuit is closed, the current in the circuit increases from zero to a certain steady value, thereby increasing the magnetic field (and hence the magnetic lines of force entering coil 2) from zero to a constant value. Similarly, when K is removed, the current, and hence the magnetic field, decreases from the steady value to zero. In both cases, the magnetic lines of force entering coil 2 change and a momentary deflection in the galvanometer is observed implying that there is a current in the galvanometer and the coil circuit.

Before discussing Faraday's law of electromagnetic induction, we shall define magnetic flux.

7.2 MAGNETIC FLUX

Similar to electric flux, the magnetic flux Φ through any surface, placed in a magnetic field **B**, is measured by the number of lines of force that cut through such a surface. Figure 7.2 shows surface S placed in a magnetic field. Consider an infinitesimal element of area ΔS on this surface. Such an element can be represented by a vector $\Delta \mathbf{S}$ whose magnitude is ΔS and whose direction is the outward drawn normal to the surface S as shown in Fig. 7.2. The magnetic flux $\Delta\Phi$ through ΔS is defined as the product of the magnitude of magnetic field **B** at a point on ΔS and the component of $\Delta \mathbf{S}$ in the direction of the magnetic field. Referring to Fig. 7.2,

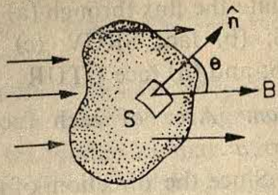


Fig. 7.2 *Surface S in a magnetic field*

$$\Delta\Phi = B\Delta S\cos\theta = \mathbf{B}\cdot\Delta\mathbf{S}$$

The total flux Φ is nothing but the sum of all such fluxes through elements over the whole surface, i.e. $\Phi = \Sigma\,\mathbf{B}\cdot\Delta\mathbf{S}$. In the limit $\Delta S \to 0$, the sum can be replaced by an integral, so that

$$\Phi = \int_S \mathbf{B}\cdot d\mathbf{S} \tag{7.1}$$

Special Cases

(i) When the surface is a plane of area A and the magnetic field is constant throughout the surface,

$$\Phi = (\mathbf{B}\cdot\hat{\mathbf{n}})A \tag{7.1a}$$

where $\hat{\mathbf{n}}$ is the unit vector perpendicular to the surface.

(ii) If the magnetic field touches the surface tangentially,

$$\theta = 90^\circ \qquad \Phi = 0 \tag{7.1b}$$

Unit of Flux

The unit of flux in the SI system is weber (Wb) and it is the flux produced when a magnetic field of 1 T (tesla) or 1 $\mathrm{Nm^{-1}\,A^{-1}}$ acts on a plane area of 1 $\mathrm{m^2}$ placed perpendicular to the direction of the magnetic field.

EXAMPLE 7.1 Consider a cube PQRSTUVW of side a placed in a uniform magnetic field **B** acting perpendicular to the face

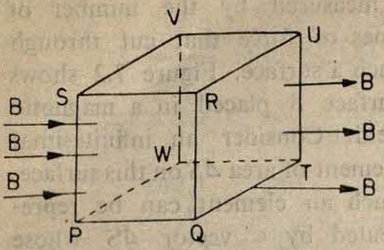


Fig. 7.3

QTUR as shown in Fig. 7.3. Calculate the flux through (a) face PQRS; (b) face SRUV; (c) face PWVS; and (d) face QTUR.

Solution: Area of each face of cube $= a^2$.

(a) Since the direction of area is perpendicular to the direction of the magnetic field so that $\theta = 90°, \Phi = 0$.

(b) Here again $\theta = 90°$, $\Phi = 0$.

(c) The outward normal from the face PWVS is opposite to **B**, so that $\theta = \pi$, and hence $\Phi = -Ba^2$.

(d) $\theta = 0, \Phi = 0$.

7.3 LAWS OF ELECTROMAGNETIC INDUCTION

The magnitude and direction of induced emf can be determined by the application of two laws of electromagnetic induction: (i) Faraday's law, and (ii) Lenz's law. The former determines the magnitude of induced emf while the latter gives the direction to the induced current. These will be discussed in detail in the following paragraphs.

Faraday's Law of Electromagnetic Induction

On the basis of experiments, Faraday found that the magnitude of the induced emf $\mathcal{E}$ was directly proportional of the rate of change of flux $\mathrm{d}\Phi/\mathrm{d}t$ associated or linked with the circuit. Mathematically, this law can be written as

$$\mathcal{E} \propto \frac{\mathrm{d}\Phi}{\mathrm{d}t}$$

Lenz's Law

The direction of induced current was first deduced by H.F. Lenz and is given by Lenz's Law. It states that whenever an induced emf is set up, the direction of induced current is such that it always opposes the cause that has produced it. This law implies that the induced emf produces such effects so as to oppose the very cause that has produced it. For example, in the first demonstration experiment described in Fig. 7.1, the cause that produced induced emf, and hence induced current in the circuit, is the motion of the magnet towards the wire loop; the direction of the induced current will be such that it (wire loop) applies a force that opposes the magnet. This can be explained as follows.

Suppose that the induced current set up in the wire is anticlockwise as seen from the magnet side towards the loop (Fig. 7.1a). This current produces a magnetic field, the direction of which can be determined from the right-hand rule. Applying this rule, one finds that the magnetic field emerges out from the loop towards the magnet, i.e. the right-hand side of the loop becomes the N-pole; and the two north poles (one due to the current loop and the other due to the bar magnet) will repel, thus opposing the motion of the magnet as predicted by Lenz's law. Similarly, if the magnet with this N-pole towards the loop is pulled away from the loop, according to Lenz's law, the induced current direction would be clockwise as seen from the magnet side, because in this case the induced current would again oppose the pull. In short, one can say that whether the magnet is moved towards or away from the coil, the induced current always opposes the motion of the magnet, as predicted by Lenz's law.

Suppose that Lenz's law is not valid and that the induced current in the loop is clockwise when the magnet (with its N-pole moving towards the loop) is pushed towards the loop (Fig. 7.1a). The magnetic field of the induced current would apply an onward force in the direction of the motion of the magnet. In such a case, one would need to move the magnet slightly and then the magnetic field of the induced current would accelerate it towards the loop, automatically increasing its kinetic energy all the time. In this case, the induced current once started would increase indefinitely with a small expenditure of energy. Such a system would be self-perpetuating and one would get enormous energy. This cannot, however, occur in nature and is not feasible considering the law of conservation of energy. Therefore, one can say that Lenz's law is valid and is a consequence of the law of conservation of energy. Further, in order to maintain the motion of the magnet, additional work has to be done against the opposing force that has arisen due to the action of the magnetic field of the induced current. This additional work produces heat in the current loop and must be equal to the joule heat produced in the circuit. If the magnet moves at a faster rate, additional work has to be done more rapidly and the rate of production of joule heat will increase correspondingly due to increase in the induced current.

Similarly, Lenz's law can also be applied to the second demonstration experiment described in Fig. 7.1b. Here the cause of the induced current in coil 2 is the change in the current in coil 1. So the induced current must be in such a direction as to oppose the current in coil 1, i.e. if the current in coil 1 is increasing, the induced current in coil 2 should produce a decreasing current in coil 1.

The procedure for finding the direction of induced current is to first assume a direction for the induced current. By applying the right-hand rule, one can find the direction of the magnetic field and its effect on the cause that produced the induced current. If the assumed direction opposes the cause, it is the right direction; otherwise, one should reverse the direction of the induced current.

So far, Lenz's law has been applied to induced currents, i.e. to closed circuits. If the circuit is open, this law can still be applied and one gets the direction of induced emf. For this purpose, one needs to imagine that the circuit is closed and then find the direction of the induced current for this imaginary circuit by applying Lenz's law. Though there is no joule heating, i.e. there is no resistive force on the magnet, yet there is an induced emf just like the emf in a battery in an open circuit.

Based on the above discussion on Lenz's law, Faraday's law can be mathematically written as

$$\mathcal{E} = -k\frac{\mathrm{d}\Phi}{\mathrm{d}t}$$

The negative sign represents the opposing action of induced current to its cause, and k is the constant of proportionality, whose value depends on the system of units chosen. In SI units, $k = 1$ and one can write

$$\mathcal{E} = -\frac{\mathrm{d}\Phi}{\mathrm{d}t} \tag{7.2}$$

If there is a coil of N turns, instead of a single turn wire loop, the total flux linked with the circuit will be N times the flux (Φ) linked with the single loop. Equation (7.2), therefore, becomes

$$\mathcal{E} = -N\frac{\mathrm{d}\Phi}{\mathrm{d}t} \tag{7.3}$$

If the flux changes uniformly ($\mathrm{d}\Phi/\mathrm{d}t$ = constant) from Φ_1 to Φ_2 in time t, Eq. (7.3) can be written as

$$\mathcal{E} = -\frac{N(\Phi_2 - \Phi_1)}{t} \tag{7.4}$$

If R is the resistance of the circuit in which induced emf is developed, the induced current $I = \mathcal{E}/R$, and the power (P) dissipated as heat is I^2R.

EXAMPLE 7.2 A rectangular coil of 100 turns and size 0.1 m × 0.05 m is placed perpendicular to magnetic field of 0.1 Wb m^{-2}. Calculate the induced emf when the magnetic field drops to 0.05 Wb m^{-2} in 0.05 s.

Solution:

$$\text{Area of the coil} = 0.1 \times 0.05\ \text{m}^2 = 5.0 \times 10^{-3}\ \text{m}^2$$

$$\text{Change in magnetic field} = 0.05 - 0.10 = -0.05\ \text{Wb m}^{-2}$$

$$\begin{aligned}\text{Change in magnetic flux} &= (\Phi_2 - \Phi_1)\\ &= -0.05 \times 5 \times 10^{-3}\ \text{Wb}\\ &= -25 \times 10^{-5}\ \text{Wb}\end{aligned}$$

$$\text{Time } t = 0.05\ \text{s}$$

Using Eq. (7.4),

$$\mathcal{E} = 100 \times \frac{25 \times 10^{-5}}{0.05} = 0.5\ \text{V}$$

EXAMPLE 7.3 An ideal solenoid of diameter 0.02 m has 600 turns per metre. At the centre of this solenoid, another coil of 100 turns is wrapped closely around it. If the current in the coil changes from 0 to 2 A in 1 millisecond, calculate the induced emf developed in the second coil.

Solution: It is known that the magnetic field near the axis of an ideal solenoid due to a current I is given by

$$B = \mu_0 n I$$

where n is the number of turns per metre and μ_0 is the permeability of vacuum.

Since the second coil is wrapped closely around the solenoid, its diameter can be taken to be equal to that of the solenoid, i.e. 0.02 m. Since the initial current and hence the initial magnetic field is zero, the change in flux for a single turn is given by

$$\mu_0 n I \pi d^2/4$$

Now, $I = 2$ A, $n = 600$, $d = 0.02$ m; substituting these values we get,

$$\text{Change in flux} = 4.74 \times 10^{-7}\ \text{Wb}$$

$$\text{Rate of change of flux} = \frac{4.74 \times 10^{-7}}{10^{-3}} = 4.74 \times 10^{-4}\ \text{Wb s}^{-1}$$

$$\text{Induced emf} \qquad \mathcal{E} = -100 \times 4.74 \times 10^{-4} = -47.4\ \text{mV}$$

7.4 ELECTROMAGNETIC INDUCTION AND LORENTZ FORCE: MOTIONAL EMF

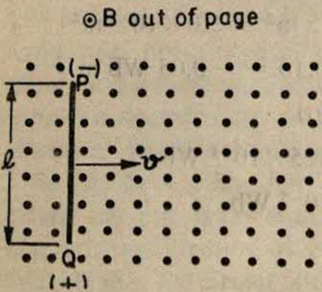


Fig. 7.4 *Induced emf in a wire moving in a magnetic field*

We can get a clear understanding of the phenomenon of electromagnetic induction if we recall that a charge q moving with a velocity $\mathbf{v}$ in a magnetic field $\mathbf{B}$ experiences a magnetic Lorentz force (see Sec. 5.5)

$$\mathbf{F} = q(\mathbf{v} \times \mathbf{B})$$

If the three vectors $\mathbf{F}$, $\mathbf{v}$, and $\mathbf{B}$ are mutually perpendicular to one another, $\mathbf{v} \times \mathbf{B} = vB$. The magnitude of $\mathbf{F}$ is qvB and it acts in a direction perpendicular to the plane of vectors $\mathbf{v}$ and $\mathbf{B}$, given by the right-hand rule.

Suppose that a wire PQ of length l moves perpendicular to the uniform magnetic field $\mathbf{B}$ as shown in Fig. 7.4. In the figure, the magnetic field is acting perpendicular to the page and out of it. According to the free electron theory of metals, the wire has free electrons in it. Consider one such electron of charge q. As the wire moves, the electron also moves with the same velocity $\mathbf{v}$; hence, it experiences a force in the upward direction towards P (because the electron has negative charge and its force will be opposite to that of the positive charge). Hence, the free electrons in the conductor move upward leaving a deficiency of electrons at Q and accumulating at P. Therefore, the end P becomes negative and the end Q becomes positive. Due to the motion of electrons an electric field is set up in the wire from P and Q whose magnitude is $F/q = vB$. As electric field is the force on a unit positive charge the work done in moving a unit positive charge from P to Q is $El = Bvl$. Hence, the induced emf at the end of the wire is Bvl.

Note: It is clear that, when a certain amount of accumulation of charges has taken place such that the electric force and the Lorentz magnetic force cancel each other, the movement of the free electrons will come to a stop. The two ends of the wire thus become like the two terminals of a cell without any external connection. This is physically the same as the Hall voltage described in Sec. 5.11.

7.5 METHODS OF GENERATING INDUCED EMF

As the induced emf is proportional to the rate of change of flux, the problem of generating induced emf reduces to the problem of changing flux. From Eq. (7.1a),

$$\Phi = BA\cos\theta = (\mathbf{B}\cdot\hat{\mathbf{n}})A$$

The magnetic flux can be changed by varying:

1. the magnitude of magnetic field **B**;
2. the magnitude of area A; and
3. the angle θ between the direction of magnetic field **B** and the direction of area A, i.e. changing the relative orientation of magnetic field and area.

1. Changing the Magnetic Field

The two demonstration experiments shown in Fig. 7.1 utilize this fact of changing the magnitude of magnetic field as explained earlier. In the first experiment (Fig. 7.1a), the motion of the magnet changes the magnetic field at any point on the wire loop while in the second experiment (Fig. 7.1b), the changing current in coil 1 (when key K is removed or inserted) changes the magnetic field at points on coil 2. In both cases, one gets induced emf.

2. Changing the Area (A) of the Coil Linked with the Magnetic Field

Consider a rectangular wire PQRS (Fig. 7.5) of length l and breadth b placed partly in a magnetic field **B** acting perpendicular and out of the page as shown. $P_1Q_1R_1S_1$ shows the limits of the magnetic field. If the wire is moved horizontally with velocity **v** perpendicular to the

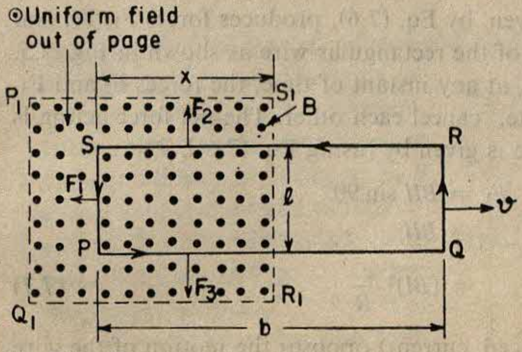


Fig. 7.5 *A rectangular wire loop moving in a magnetic field*

side QR, the area of the rectangle in the region of the magnetic field changes, thereby changing the magnetic flux linked with the circuit. Hence, the induced emf is set up in the wire. This emf can be calculated as follows.

Let x be the length of the wire in the magnetic field at any instant of time t. The flux linked with the circuit at this instant is Blx, for lx is the area of wire in the magnetic field. In a further time interval Δt, let Δx be the change in length. Then

$$\text{Change in area} = l\Delta x$$

$$\text{Change in flux} = Bl\Delta x$$

$$\text{Rate of change of flux} = -Bl\frac{\Delta x}{\Delta t}$$

Hence,

$$\text{Induced emf} = -Bl\frac{\Delta x}{\Delta t}$$

But, $\Delta x/dt = -v$, because as the wire moves to the right, x decreases with time, i.e. $\Delta x/\Delta t$ is negative. Also, the magnitude of **v** is $\Delta x/\Delta t$.

$$\therefore \qquad \mathcal{E} = Blv \tag{7.5}$$

which is the same as the emf set up at the ends of the wire moving with velocity **v**.

If R is the resistance of the circuit, the induced current I is

$$I = \frac{\mathcal{E}}{R} = \frac{Blv}{R} \tag{7.6}$$

It is known that a wire of length l, carrying a current I, moving in a magnetic field **B** experiences a force **F** given by

$$\mathbf{F} = I(\mathbf{l} \times \mathbf{B})$$

The induced current, given by Eq. (7.6), produces forces $\mathbf{F}_1$, $\mathbf{F}_2$, and $\mathbf{F}_3$ on the three segments of the rectangular wire as shown in Fig. 7.5. It can be easily seen that, at any instant of time, the forces $\mathbf{F}_2$ and $\mathbf{F}_3$, being equal and opposite, cancel each other. The net force acting is only $\mathbf{F}_1$, whose magnitude is given by [using Eq. (7.6)].

$$\begin{aligned} F_1 &= BIl \sin 90^\circ \\ &= BIl \\ &= (Bl)^2 \frac{v}{R} \end{aligned} \tag{7.7}$$

This force (due to induced current) opposes the motion of the wire (from Lenz's law) and hence to maintain the motion of the wire with the same velocity, one has to do work at a steady rate P given by

$$P = F_1 v = \frac{(Blv)^2}{R} \tag{7.8}$$

P is nothing but the power supplied to maintain the motion of the wire and one can see [using Eq. (7.6)] that P is equal to I^2R, i.e. rate of joule heat production.

3. Changing the Relative Orientation of Magnetic Field and Area of Coil

Consider a rectangular coil ABCD of N turns and area A placed in a uniform magnetic field **B** (Fig. 7.6). This coil rotates in an anticlockwise direction with an angular velocity ω about an axis perpendicular to the direction of the magnetic field. Suppose, initially the coil is vertical so that the angle between normal to the plane of the coil and magnetic field is zero. At a later time t, let θ be the angle through which it has rotated so that $\theta = \omega t$. At this instant, the flux $\Phi = NAB \cos\theta$ and hence the induced emf $\mathcal{E} = -\mathrm{d}\Phi/\mathrm{d}t = +NAB \sin\theta \, \mathrm{d}\theta/\mathrm{d}t = NAB\omega \sin\theta$. Hence ($\because \theta = \omega t$)

$$\mathcal{E} = NAB\omega \sin \omega t \tag{7.9}$$

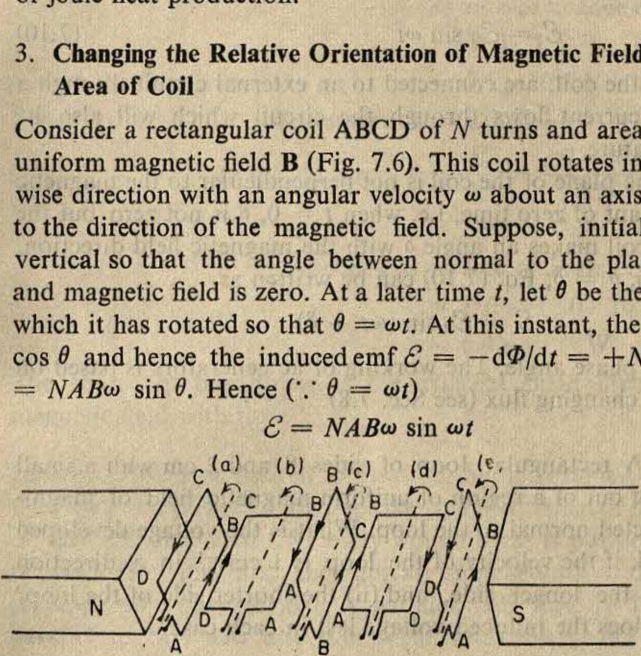


Fig. 7.6 *The rotation of the coil in a uniform magnetic field: a, b, c, d, and e are the different positions of the coils when $\theta = 0, \pi/2, \pi, 3\pi/2$, and 2π, respectively*

Equation (7.9) gives the induced emf $\mathcal{E}$ at any instant of time t as the coil rotates; $\mathcal{E}$ is called instantaneous emf induced in a rotating coil. If the coil is rotating with constant angular velocity ω, then N, A, B and ω are constant and the induced emf $\mathcal{E}$ varies sinusoidally with time; as shown in Fig. 7.7. When the coil is vertical, i.e. $\theta = 0$ or $180°$, $\mathcal{E} = 0$ as can be seen from Eq. (7.9). When the coil is in the horizontal position, i.e. $\theta = \pi/2$

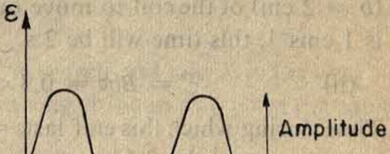


Fig. 7.7 *Variation of $\mathcal{E}$ with t*

or $3\pi/2$, the magnitude of the emf is maximum but the sign of emf is reversed. So, the emf alternates, reversing the direction twice in each rotation. The maximum emf $\mathcal{E}_0 = NAB\omega$ is called the amplitude of emf. Further, as the coil completes one full revolution, i.e. θ varies from 0° to 360°, the induced emf at the ends completes one cycle. The induced emf can be written as

$$\mathcal{E} = \mathcal{E}_0 \sin \omega t \tag{7.10}$$

If the ends of the coil, are connected to an external circuit through a resistance R, a current flows through the circuit which will also be sinusoidal in nature.

Further, if the plane of the coil is not perpendicular to the magnetic field at the instant of zero time, i.e. when $t = 0$, θ is not zero but the normal to the coil makes an angle δ with the magnetic field direction, i.e. when $t = 0$, $\theta = \delta$, Eq. (7.10) can be written as

$$\mathcal{E} = \mathcal{E}_0 \sin (\omega t + \delta)$$

δ is called the phase angle. The working of ac generators is based on this method of changing flux (see Sec. 7.8).

EXAMPLE 7.4 A rectangular loop of sides 8 and 2 cm with a small break is moving out of a region of uniform magnetic field of magnitude 0.3 T directed normal to the loop. What is the voltage developed across the break if the velocity of the loop is 1 cms^{-1} in a direction normal to: (i) the longer side, and (ii) the shorter side of the loop? For how long does the induced voltage last in each case?

Solution: $B = 0.3$ T $\quad v = 1$ cms$^{-1} = 0.01$ ms^{-1}

$l = 8$ cm $= 0.08$ m $\quad b = 2$ cm $= 0.02$ m

(i) $\mathcal{E} = Blv = 0.3 \times 0.08 \times 0.01 = 2.4 \times 10^{-4}$ T

The time during which the emf lasts is the time taken by the breadth ($b = 2$ cm) of the coil to move out of the field. Since the speed of coil is 1 cms^{-1}, this time will be 2 s.

(ii) $\mathcal{E} = Bbv = 0.3 \times 0.02 \times 0.01 = 0.6 \times 10^{-4}$ T

Time during which this emf lasts = time taken by the length ($l = 8$ cm) to move out of the field = 8 s.

EXAMPLE 7.5 Suppose the loop in Example 7.4 above is stationary but the current feeding the electromagnet that produces the magnetic field is gradually reduced so that the field decreases from its initial value of 0.3 T at the rate of 0.02 T s^{-1}. If the break is joined and the loop has a resistance of 1.6 Ω how much power is dissipated by the loop as heat? What is the source of this power?

Solution:

$$\begin{aligned}
\text{Area of the coil} &= 8\times 2\times 10^{-4}\ \text{m}^2\\
\text{Rate of change of magnetic field} &= 0.02\ \text{T s}^{-1}\\
\therefore\quad \text{Induced emf } (\mathcal{E}) &= \text{rate of change of flux}\\
&= \text{rate of change of } B\times\text{area}\\
&= (0.02\ \text{T s}^{-1})\times(8\times 2\times 10^{-4}\ \text{m}^2)\\
&= 3.2\times 10^{-5}\ \text{V}
\end{aligned}$$

Now

$$\begin{aligned}
\text{Resistance } R &= 1.6\ \Omega\\
\therefore\quad \text{Induced current } I &= \frac{\mathcal{E}}{R} = \frac{3.2\times 10^{-5}}{1.6}\\
&= 2.0\times 10^{-5}\ \text{A}\\
\text{Power loss} &= \text{rate of Joule heating}\\
&= I^2R = (2.0\times 10^{-5})^2\times 1.6\\
&= 6.4\times 10^{-10}\ \text{W}
\end{aligned}$$

The source of this power is the external agent used to change the magnetic field with time.

EXAMPLE 7.6 It is desired to measure the magnitude of field between the poles of a powerful loud-speaker magnet. A small flat search coil of area 2.0 cm^2 with 25 closely wound turns is positioned normal to the field direction, and then quickly snatched out of the field region. (Equivalently, one can give it a quick 90° turn to bring its plane parallel to the field direction.) The total charge flowing in the coil (measured by a ballistic galvanometer connected to the coil) is 7.5 mC. The resistance of the coil and the galvanometer is 0.50 Ω. Estimate the field strength of the magnet.

Solution: The induced current is given by [see Eq. (7.3)]

$$I = \frac{\mathcal{E}}{R} = -\frac{N}{R}\frac{\mathrm{d}\Phi}{\mathrm{d}t}$$

where N is the number of turns in the search coil and R is the total resistance in the circuit. The charge flowing through any point in the circuit is given by $\left(I = \frac{\mathrm{d}Q}{dt}\right)$

$$\begin{aligned}
Q = \int_{t_\mathrm{i}}^{t_\mathrm{f}} I\,\mathrm{d}t &= -\frac{N}{R}\int_{t_\mathrm{i}}^{t_\mathrm{f}} \mathrm{d}\Phi = -\frac{N}{R}(\Phi_\mathrm{f}-\Phi_\mathrm{i})\\
&= \frac{N}{R}(\Phi_\mathrm{i}-\Phi_\mathrm{f}) \qquad \text{(i)}
\end{aligned}$$

If A is the area of the coil, then the initial flux (i.e. flux when the search coil is normal to field **B**) is

$$\Phi_i = BA$$

The final flux (when the coil is removed from the field or when it is parallel to **B**) is obviously zero, i.e.

$$\Phi_f = 0$$

so that

$$\Phi_i - \Phi_f = BA \qquad \text{(ii)}$$

Using (ii) in (i) gives

$$Q = \frac{NBA}{R}$$

or

$$B = \frac{QR}{NA} \qquad \text{(iii)}$$

We are given that $Q = 7.5$ mC $= 7.5 \times 10^{-3}$ C, $R = 0.5\ \Omega$, $N = 25$ and $A = 2.0$ cm^2 $= 2.0 \times 10^{-4}$ m^2. Substituting these values in (iii) and solving, we get

$$B = 0.75 \text{ T}$$

EXAMPLE 7.7 Figure 7.8 shows a metal rod PQ resting on the rails AB and positioned between the poles of a permanent magnet. The rails, the rod and the magnetic field are in three mutually perpendicular directions. A galvanometer G connects the rails through a switch K. Length of the rod = 15 cm, $B = 0.50$ T, resistance of the closed loop containing the rod = 9.0 mΩ.

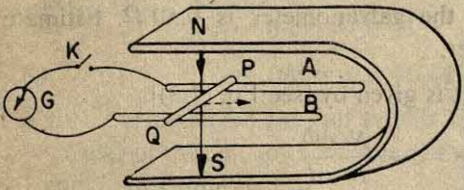


Fig. 7.8

Answer the following questions:

(a) Suppose K is open and the rod moves with a speed of 12 cms^{-1} in the direction shown. Give the polarity and magnitude of the induced emf.

(b) Is there an excess charge built up at the ends of the rods when K is open? What if K is closed?

(c) With K open and the rod moving uniformly, there is no net force on the electrons in the rod PQ even though they do experience magnetic force due to the motion of the rod. Explain.

(d) What is the retarding force on the rod when K is closed?

(e) How much power is required (by an external agent) to keep the rod moving at the same speed (= 12 cms^{-1}) when K is closed?

(f) How much power is dissipated as heat in the closed circuit? What is the source of this power?

(g) What is the induced emf in the moving rod when the permanent magnet is rotated to a vertical position so that the field is parallel to the rails?

Solution: (a) The magnitude of the induced emf is

$$\mathcal{E} = vBl \sin \theta$$

where θ is the angle between the axis of the rod and the magnetic field, which in this case is 90°. Hence

$$\begin{aligned}\mathcal{E} &= vBl \\ &= (12\times10^{-2})\times0.5\times(15\times10^{-2}) \\ &= 9.0\times10^{-3}\ \text{V}\end{aligned}$$

It is clear from the diagram that the electrons will move toward end Q (which acquires a negative charge) and the end P will acquire a positive charge (due to a deficiency of electrons). Hence P is the positive end and Q the negative end.

(b) Yes, because emf is induced across the ends of the rod due to its motion in the field. When key K is closed, the induced current maintains the excess charge.

(c) The excess charge on the ends of the rod sets up an electric field E. The force $e\mathbf{E}$ due to electric field is cancelled by the Lorentz magnetic force $e(\mathbf{v}\times\mathbf{B})$. Here e is the magnitude of the charge on an electron.

(d) Resistance = 9.0 mΩ = $9.0\times10^{-3}\ \Omega$. The induced current is

$$I = \frac{\mathcal{E}}{R} = \frac{9.0\times10^{-3}\ \text{V}}{9.0\times10^{-3}\ \Omega} = 1.0\ \text{A}$$

$$\begin{aligned}\therefore \qquad \text{Retarding force } F \text{ on the rod} &= BIl \\ &= 0.5\times1.0\times0.15 \\ &= 7.5\times10^{-2}\ \text{N}\end{aligned}$$

(e) Power required by an external agent to keep the rod in uniform motion with a velocity $v = 12\times10^{-2}\ \text{ms}^{-1}$ against the retarding force F is

$$P = \text{force} \times \text{velocity} = Fv$$

$$= (7.5\times10^{-2})\times(12\times10^{-2}) = 9.0\times10^{-3}\ \text{W}$$

(f) Power dissipated as heat $= I^2R$

$$= (1.0)^2\times9.0\times10^{-3}$$

$$= 9.0\times10^{-3}\ \text{W}$$

The source of power is the external agent which keeps the rod in motion.

(g) In this case angle $\theta = 0$. Hence $\mathcal{E} = vBl \sin\theta = 0$. The reason is that the moving rod does not cut across the field lines, and therefore there is no flux change. Consequently, the induced emf is zero.

EXAMPLE 7.8 A rectangular coil 0.20 m × 0.10 m and having 100 turns rotates in a magnetic field of induction 0.003 Wb m^{-2} with a frequency of 1200 rpm about an axis normal to the magnetic field.

What is:

(a) the maximum value of induced emf?

(b) the instantaneous emf when the plane of the coil is perpendicular to the magnetic field?

(c) the instantaneous emf when the plane of the coil is making an angle of 30° with the magnetic field?

Solution:

$$\text{Area of coil } A = 0.02\ \text{m}^2$$

$$B = 0.003\ \text{Wb m}^{-2}$$

$$N = 100 \qquad \omega = 2\pi\times20\ \text{rps} = 40\pi\ \text{rad s}^{-1}$$

(a) $\mathcal{E}_0 = NA\omega B = 100\times0.02\times0.3\times10^{-2}\times40\pi = 0.75\ \text{V}$

(b) Zero, as $\theta = 0$.

(c) As the plane of the coil is making an angle of 30° with the magnetic field the angle, $\theta = 90 - 30 = 60°$

$$\therefore \quad \mathcal{E} = \mathcal{E}_0 \sin 60° = 0.75 \sin 60° = 0.65\ \text{V}$$

7.6 EDDY CURRENTS

So far, the induced currents in a circuit due to induced emf, generated in a wire or ring of definite geometry have been considered. It was observed first by Foucault that when a solid mass of any conducting substance moves in a non-uniform magnetic field, induced currents are set up in the solid mass in the form of eddies. Such currents are known as *eddy currents* or *Foucault currents*. These currents (moving charges) cause a magnetic Lorentz force to be exerted on them. The following two experiments demonstrate eddy currents and the force due to them.

1. Take an electromagnet fed by an ac source and place a light metallic disc on its top face as shown in Fig. 7.9.

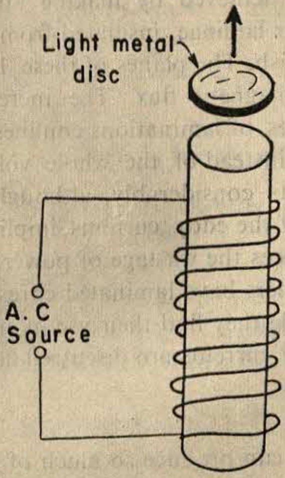


Fig. 7.9 *A light metallic disc placed on top of an electromagnet is thrown up when the current is switched on*

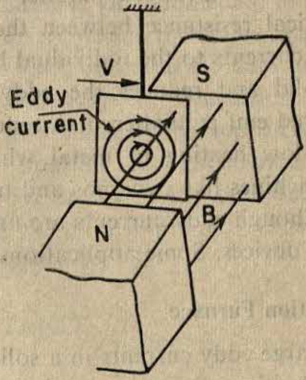


Fig. 7.10 *Eddy currents damp the swinging solid mass in a magnetic field*

As the current is switched on, the magnetic field at the disc rises from zero to a finite value, causing induced currents that effectively convert it into a small magnet. If initially the top end of the electromagnet starts acquiring N-polarity, by Lenz's rule the small magnetic disc will have its N-pole next to it, resulting in a repulsive force. The light metallic disc is thus seen to be thrown up as the current in the electromagnet is switched on.

2. When a conductor moves in a magnetic field, eddy currents in it cause a force on it, which, by Lenz's law, opposes the motion.

Figure 7.10 shows a copper pendulum, i.e. a solid sheet of copper suspended from a fixed end between the poles NS of the magnet. The direction of these currents can be found by applying Lenz's law. These currents tend to set up a magnetic field which opposes the motion of the solid mass with such a strong retarding force that the whole mass almost comes to rest because the resistance of the solid mass is very low and the eddy currents are very high. In such a case, the solid body behaves as though it is embedded in a very viscous medium.

These eddy currents produce heat and hence consume a lot of energy. This is undesirable in a number of devices like transformers, dynamos, etc. where the coil is wound on iron cores. These currents can be minimized, although not completely eliminated, by laminating the solid mass of iron or metal. This is achieved by making up the solid mass of a bundle of thin sheets or laminae, insulated from one another by the use of an insulating varnish. The planes of these laminations are arranged parallel to the magnetic flux. The increased electrical resistance between the surfaces of laminations confines the eddy currents to the individual laminae instead of the whole volume of solid and reduces the eddy currents considerably, although the induced emf is unaltered. A reduction of the eddy currents implies a very low heating of metal, which reduces the wastage of power. All ac machines like dynamos and transformers have laminated cores.

Although eddy currents are undesirable they find their use in many other devices. Some applications of eddy currents are discussed here.

Induction Furnace

The large eddy currents in a solid metal can produce so much of heat that melting of metal can take place. For this, the substance to be heated is placed in a rapidly changing or high frequency magnetic field which produces large eddy currents in the substance. This, in turn, heats the substance so much that it melts. This is the principle of an induction furnace where a metal is separated from its ore or some alloys are made (which otherwise cannot be prepared).

This heating effect of eddy currents can also be used to heat localized tissues of a patient.

Electric Brakes

Electric brakes or eddy current brakes are used in stopping electric trains. The principle of these brakes is that when a magnetic field is suddenly applied to a rotating metallic disc, eddy currents are set up in the disc which in turn exert a torque on the disc so as to stop it.

Galvanometer Damping

The application of eddy currents can also be made in the construction of moving coil galvanometers. When a steady current is passed in a galvanometer, the coil is deflected through a certain angle and reaches a position of equilibrium. Usually, the coil does not come to its position of equilibrium immediately but oscillates about it for some time. To bring it immediately to its position of equilibrium the frame on which the galvanometer coil is wound is made of some metal like brass or copper. As soon as the coil and the frame rotate in the magnetic field, eddy currents are set up in the matallic frame which in turn resist the motion of the galvanometer coil. Thus, the motion of the frame is damped and the coil attains its equilibrium position at once. This damping is known as *electromagnetic damping*. In ballistic galvanometers, this is not desirable and the frame is made of some non-metallic substance. In such cases, the oscillations of the galvanometer are stopped by short-circuiting the galvanometer coil because then large induced currents are set up in the coil due to its motion in the magnetic field. That is why, key K in Fig. 7.12 is always connected parallel to a ballistic galvanometer G.

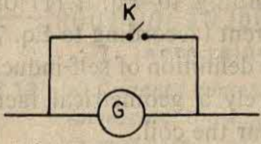


Fig. 7.11 *Galvanometer with a key*

7.7 SELF AND MUTUAL INDUCTANCE

Self-Inductance

Let us refer to Fig. 7.1(b) again. If the two coils are placed near each other a current I in one coil will set up a flux through the second coil. If this flux is changed by changing the current, an induced emf will appear in the second coil according to Faraday's law. However, two coils are not needed to show an inductive effect. An induced emf appears in a *single* coil if the current in *that* coil is changed. The current produces a magnetic field which threads every turn of the coil. If the current is changing with time, the magnetic field will also change. Consequently the magnetic flux threading the coil changes, resulting in an induced emf. Thus whenever there is a change in the current flowing in a coil, an induced emf is set up in the coil itself. This emf is called the *self-induced* emf and it obeys Faraday's law of induction like any other induced emf.

For a given coil, provided no magnetic material such as iron is nearby, the magnetic flux linked with it will be proportional to the current, i.e.

$$\Phi \propto I$$

or

$$\Phi = LI \tag{7.11}$$

where L is a constant of proportionality called the *self-inductance* (or simply *inductance*) of the coil. Using Faraday's law [Eq. 7.2)] and Eq. (7.11), the self-induced emf is given by

$$\mathcal{E} = -\frac{\mathrm{d}\Phi}{\mathrm{d}t} = -L\frac{\mathrm{d}I}{\mathrm{d}t} \tag{7.12}$$

Equations (7.11) and (7.12) define the self-inductance of a coil as the total flux linkage per unit current (according to Eq. 7.11) or the induced emf for unit rate of change of current (according to Eq. 7.12). It is worth mentioning here that though the definition of self-inductance involves induced emf and current, it is purely a geometrical factor if iron or a similar material is not present near the coil.

From Eqs (7.11) and (7.12) the unit of inductance can be *weber per ampere* or *volt-second per ampere* which, of course, are equivalent to each other. A special name called the *henry* (symbol H) is given to these combination of units, i.e.

$$1\ \mathrm{H} = 1\ \mathrm{WbA^{-1}} = 1\ \mathrm{VsA^{-1}}$$

The unit of inductance is named after Joseph Henry (1797–1878), an American physicist and a contemporary of Faraday. Henry independently discovered the law of induction at about the same time Faraday did. The henry is a big unit of inductance. Smaller units *millihenry* (mH) $= 10^{-3}$H and microhenry (μH) $= 10^{-6}$H are often used.

The self-inductance of a coil is one henry if an induced emf of one volt is set up when the current in the coil changes at the rate of one ampere per second.

The direction of the induced emf can be found from Lenz's law. Suppose a steady current I (produced by a battery) is flowing in a coil. Let us suddenly reduce the battery emf to zero (by disconnecting it). The current I will immediately begin to decrease. This decrease in current is responsible for self-induced emf $\mathcal{E}$. From Lenz's law, the induced emf must oppose this decrease in current. To oppose a falling current, the induced emf must point in the *same* direction as the current, as in Fig. 7.12(a). On the other hand, when the current in the coil is *increased*, Lenz's law shows that the self-induced emf points in the direction *opposite* to that of the current, as in Fig. 7.12(b). In each case, the self-induced emf acts to oppose the *change* in current. The minus sign in Eq. (7.12) shows that $\mathcal{E}$ and $\mathrm{d}I/\mathrm{d}t$ are opposite in sign, since L is always a positive quantity. The value of L depends on the length, number of turns and the cross-sectional area of the coil.

An ideal inductor is one that has zero resistance. It should be noted that a straight wire has self-inductance, through small, because the

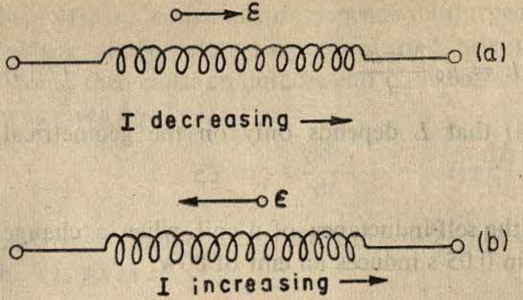


Fig. 7.12 *(a) Current I decreasing and (b) Current I increasing; the emf opposes the change in each case*

current in the wire contributes to the magnetic flux linking the circuit of which it is a part. Usually one neglects the self-inductance of a straight wire but it plays an important role in very high frequency circuits. Further, the role of the magnetic field in an inductor is extremely significant and it is similar to that of the electric field in a capacitor. From the above discussion one can say that the self-inductance in any circuit is the property of the circuit which opposes any change of current in the circuit and hence the role of inductance in any electrical circuit is the same as that of inertia in mechanics. Therefore, self-inductance is also called the measure of electrical inertia.

Self-Inductance of a Long Solenoid

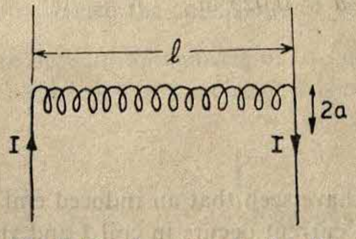


Fig. 7.13 *Calculation of self-inductance*

Consider a long and closely wound solenoid of radius a and length $l (l \gg a)$ as shown in Fig. 7.13. Let N be the total number of turns in the solenoid. The magnetic induction on the axis of a long solenoid due to a current I is given by

$$B = \mu_0 \frac{NI}{l}$$

$$\text{Magnectic flux per turn} = \mu_0 \frac{\pi a^2 I}{l}$$

Total magnetic flux linked with the solenoid of N turns,

$$= \mu_0 \frac{N}{l} \pi a^2 I N$$

$$= \frac{\mu_0 \pi a^2 N^2 I}{l}$$

Since, $\Phi = LI$, we have

$$L = \mu_0 \frac{N^2 \pi a^2}{l} \tag{7.14}$$

It is clear from Eq. (7.14) that L depends only on the geometrical factors of the solenoid.

EXAMPLE 7.9 Calculate the self-inductance of a coil when a change of current from 0 to 2A in 0.05 s induces an emf of 80 V.

Solution:

$$\text{Change in current} = 2\text{A}$$

$$\text{Time} = 0.05 \text{ s}$$

$$\text{Rate of change of current} = \frac{2}{0.05} = 40 \text{ As}^{-1}$$

$$\text{Self-induced emf} = 80 \text{ V}$$

$$\text{Self-inductance } L = \frac{80}{40} = 2 \text{ H}$$

EXAMPLE 7.10 What is the self-inductance of an air core solenoid 1 m long, diameter 0.05 m, if it has 700 turns?

Solution:

$$L = \frac{\mu_0 \pi a^2 N^2}{l}$$

Here $\mu_0 = 4\pi \times 10^{-7}$ Hm^{-1} $\quad a = 0.025$ m $\quad N = 700$

$l = 1$m

$\therefore \quad L = 1.21$ mH

Mutual Inductance

Let us refer to Fig. 7.1(b) again. We have seen that an induced emf is set up in coil 2 whenever a change in current occurs in coil 1 and vice versa. This process is called *mutual Induction.* If no iron material is present near the coils, the flux linkage in coil 2 depends on the magnetic field in coil 1 which in turn depends on the current I_1 in coil 1. The flux Φ_{21} through coil 2 due to current I_1 in coil 1 can, therefore, be written as

$$\Phi_{21} \propto I_1$$

or

$$\Phi_{21} = M_{21} I_1 \tag{7.15}$$

where M_{21} is a constant that depends on the geometry of the coils, i.e. their size, shape, relative separation and orientation. Any change in I_1 would then cause an induced emf $\mathcal{E}_2$ in coil 2, which from Faraday's law is given by

$$\mathcal{E}_2 = -\frac{\mathrm{d}\Phi_{21}}{\mathrm{d}t} = -M_{12}\frac{\mathrm{d}I_1}{\mathrm{d}t} \tag{7.16}$$

Similarly, a change in current I_2 in coil 2 induces an emf $\mathcal{E}_1$ in coil 1 which is given by

$$\mathcal{E}_1 = -M_{12}\frac{\mathrm{d}I_2}{\mathrm{d}t}$$

We will show below that the constants M_{21} and M_{12} are equal to each other; $M_{21} = M_{12} = M$.

Here M is called the mutual inductance of the coils. It is the property of a pair of circuits by virtue of which any change of current in one circuit induces an emf in the other circuit. In addition to the current, the magnetic flux linkage also depends on the geometrical factors as well as the relative placement of two coils. If two coils are placed some distance apart, the magnetic flux linkage will be less for the same current and hence M will also be low. If the coils are very tightly wound one over another, most of the flux produced in the primary will be linked with the secondary for the same current and M will be more (see Fig. 7.14). In other words, one can say that the coupling between the two coils is more in the latter case than in the former. As a quantitative measure of this coupling one defines a quantity called the coefficient of coupling $k = \sqrt{\frac{M}{L_1L_2}}$ where L_1 and L_2 are the self-inductances of two coils; k is always less than one. For better efficiency of the circuit k is made as large as possible.

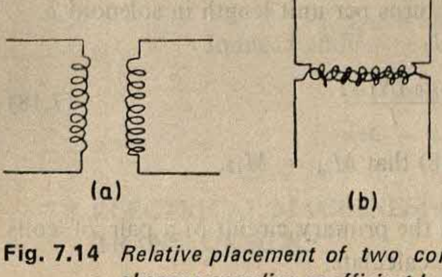

(a) (b)

Fig. 7.14 *Relative placement of two coils changes coupling coefficient k, in (a) k is small while in (b) k is large*

As in self-inductance, the mutual inductance between two coils may be defined mathematically as the emf induced in the secondary coil for unit rate of change of current in the primary.

The unit of mutual inductance is the henry, which is the mutual inductance of that pair of circuits in which a rate of change of current of one ampere per second in one circuit induces an emf of one volt in the second circuit.

Calculation of Mutual Inductance

Consider two long solenoids of length l, one having N_1 turns and carrying a current I_1, completely surrounding the other one which has N_2 turns and carries a current I_2 [see Fig. 7.14(b)]. The coils of one are closely wound over those of the other so that they have a common cross-sectional area, say, $A = \pi a^2$, where a is the common radius. The axes of the two coils are kept parallel. The fields due to the solenoids can be considered uniform inside them and zero outside. The flux linked with the inner solenoid (called solenoid 2) is

$$\Phi_{21} = \text{magnetic field} \times \text{area} \times \text{number of turns}$$
$$= (\mu_0 n_1 I_1) \times A \times N_2$$

where $n_1 = N_1/l$ is the number of turns per unit length in solenoid 1. Hence from Eq. (7.15), we have

$$M_{21} = \mu_0 n_1 A N_2 = \frac{\mu_0 A N_1 N_2}{l} \tag{7.17}$$

Consider now the flux linked with the outer solenoid (called solenoid 1) due to current I_2 in solenoid 2. The field due to I_2 is constant inside the inner solenoid but zero in the annular region between the two solenoids. Hence

$$\Phi_{12} = \mu_0 n_2 I_2 A N_1$$

where $n_2 = N_2/l$, the number of turns per unit length in solenoid 2. Hence

$$M_{12} = \mu_0 n_2 A N_1 = \frac{\mu_0 A N_1 N_2}{l} \tag{7.18}$$

Notice from Eqs (7.17) and (7.18) that $M_{21} = M_{12}$.

EXAMPLE 7.11 If the current in the primary circuit of a pair of coils changes from 10 A to 0 in 0.1 s, calculate:

(a) the induced emf in the secondary if the mutual inductance between the coils is 2 H; and

(b) the change of the flux per turn in the secondary if it has 500 turns.

Solution:

(a) $$\mathcal{E}_s = -M\frac{dI}{dt} = +2\times\frac{10}{0.1} = 200\text{ V}$$

Note that $\frac{dI}{dt}$ is negative.

(b) Also,

$$\mathcal{E}_s = -N_s\frac{d\Phi}{dt}$$

where N_s is the number of turns in the secondary. Given $N_s = 500$.

$$\therefore \quad d\Phi = \frac{200}{500}\times 0.1 = 0.04\text{ Wb}$$

Change of flux per turn = 0.04 Wb

EXAMPLE 7.12 A solenoid 1.0 m long and 0.05 m in diameter, has 700 turns. Another solenoid of 50 turns is tightly wound over the first solenoid. Find: (a) the mutual inductance of the two solenoids, and (b) the induced emf in the second solenoid, when the current in the first changes from 0 to 5.0 A in 0.01s.

Solution:

(a) $$M = \frac{\mu_0 A N_1 N_2}{l} \qquad \text{(i)}$$

Here $l = 1.0$ m, $A = \pi(0.025)^2\text{ m}^2$, $N_1 = 700$, $N_2 = 50$, and $\mu_0 = 4\pi \times 10^{-7}\text{ Hm}^{-1}$. Using these values in (i), we get

$$M = 8.6\times 10^{-5}\text{ H}$$

(b) $$\frac{dI}{dt} = \frac{5.0}{0.01} = 500\text{ A s}^{-1}$$

$$\begin{aligned}\therefore \quad \text{Induced emf} &= -M\frac{dI}{dt}\\ &= -8.6\times 10^{-5}\times 500\\ &= -4.3\times 10^{-2}\text{ V}\end{aligned}$$

7.8 ELECTRICAL MACHINES: AC GENERATOR AND TRANSFORMER

The phenomenon of electromagnetic induction is the basis of many modern electrical devices, such as generators and transformers.

The Alternating Current Generator

The alternating current generator which is one of the most important applications of the phenomenon of electromagnetic induction converts mechanical energy into electrical energy. The modern form of the generator is due to Nikola Tesla, the Yugoslav scientist and inventor.

A rectangular coil is placed between the pole pieces of a large permanent magnet or electromagnet. The shape of the pole pieces is usually flat, unlike in the case of a galvanometer (Fig. 7.15). When the coil rotates by supplying mechanical energy, an induced emf is developed at its ends. If the coil is wound on an iron core the flux linked with the coil increases (for the permeability of iron and, hence, magnetic induction is much more than that in case of wood, air, etc.) and a large emf may be obtained. As the emf alternates, the generator is called an alternating current (ac) generator or dynamo and the coil is called the armature. The ends of the coil are connected to two circular rings C_1 and C_2 called the slip or connecting rings. The connections of the external circuit are made with the help of two stationary carbon (graphite) brushes B_1 and B_2 which are permanently in contact with C_1 and C_2, as shown in Fig. 7.15. Basically, the ac generator has three components: a field magnet, an armature, and slip rings with brushes. The emf generated in this way can be used to run machines, to provide electricity in houses, etc.

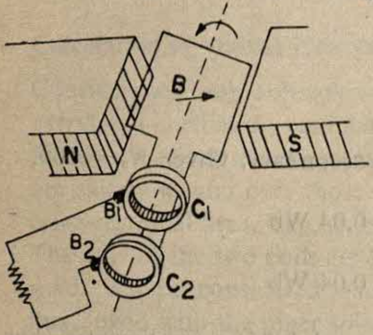


Fig. 7.15 *A schematic diagram of an ac generator*

When the plane of the coil is normal to the magnetic field **B** the flux through it is NAB, where N is the number of turns and A is the area of each turn. If the coil rotates with an angular speed ω, then let θ be the angle between the normal to the coil and magnetic field **B** at a certain instant of time t. The flux at time t is, $\theta = \omega t$ since

$$\Phi = NAB \cos \omega t$$

From Faraday's law, the induced emf $\mathcal{E}$ is given by

$$\mathcal{E} = -\frac{\mathrm{d}\Phi}{\mathrm{d}t} = -\frac{\mathrm{d}}{\mathrm{d}t}(\mathrm{NAB} \cos \omega t)$$

$$= \mathrm{NAB}\, \omega \sin \omega t$$

or

$$\mathcal{E} = \mathcal{E}_0 \sin \omega t \qquad (7.19)$$

where $\mathcal{E}_0 = NAB\omega$. If the ends of the coil are well outside the field and if the coil is on an open circuit, i.e. the ends are not connected to an external circuit, then we can think of the ends as terminals having a potential difference $\mathcal{E}$. When a load (resistor R) is connected across the terminals, a current I flows through the circuit.

$$I = \frac{\mathcal{E}}{R} = \frac{\mathcal{E}_0}{R} \sin \omega t = I_0 \sin \omega t \tag{7.20}$$

where $I_0 = \mathcal{E}_0/R$.

The current I and voltage $\mathcal{E}$ both vary sinusoidally with time. Such a current is called ac or alternating current.

The principle of the generator outlined above is the basis for the working of practical devices for generating electrical power, ranging from small portable generators to giant thermal and hydroelectric power stations. In a hydroelectric power station water is stored in a dam at a height from where it falls onto giant turbines. These turbines are connected to the loops of wires in ac generators. The kinetic energy of the falling water thus gets converted into rotational energy of the turbines and ultimately into electrical energy supplied by the generator. In a thermal power station, the turbines are rotated by super heated steam gushing past them. Steam is produced by boiling water using coal or oil as fuel.

The Transformer

Another common electrical device whose working is also based on the principle of electromagnetic induction is the transformer. It is a device used for converting a low voltage to a high voltage and vice versa. The former is called the step-up transformer and the latter the step-down transformer.

A transformer consists of two coils each of which is wound on an iron core as shown in Fig. 7.16. One of the coils is connected to a source of alternating emf. This coil is called the *primary* of the transformer while the other is called the *secondary* of the transformer. Any of the two coils can act as primary while the other as secondary. The alternating emf in one coil induces an alternating emf in the second coil. The presence of an iron core in the primary and secondary makes the flux linkage between the two coils very large. The alternating emf in the coil makes the magnetic flux in the iron also vary periodically. This varying magnetic flux in iron induces an alternating emf in the secondary.

If the magnetic field lines remain confined to the core, then all the field lines threading the primary also go across the secondary. Then

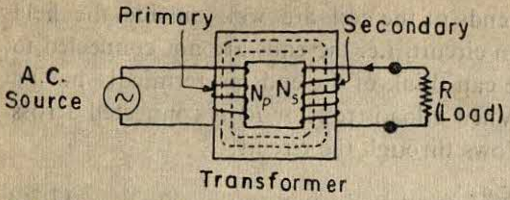


Fig. 7.16 *A schematic diagram of a transformer. Flux linkage is shown by broken lines*

the magnetic fluxes across the secondary and primary will be simply proportional to the number of turns in them, i.e.

$$\frac{\Phi_s}{\Phi_p} = \frac{N_s}{N_p}$$

or

$$\Phi_s = \left(\frac{N_s}{N_p}\right)\Phi_p$$

where N_s = number of turns in the secondary and N_p = number of turns in the primary. Now from Faraday's law the emf induced across the secondary is $\mathcal{E}_s = -(\mathrm{d}\Phi_s/\mathrm{d}t)$ and that across the primary is $\mathcal{E}_p = -(\mathrm{d}\Phi_p/\mathrm{d}t)$. Thus

$$\mathcal{E}_s = -\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{N_s}{N_p}\cdot\Phi_p\right)$$

$$= -\frac{N_s}{N_p}\frac{\mathrm{d}\Phi_p}{\mathrm{d}t}$$

or

$$\mathcal{E}_s = \frac{N_s}{N_p}\mathcal{E}_p \qquad (7.21)$$

From this equation, it follows that if $N_s > N_p$; $\mathcal{E}_s > \mathcal{E}_p$, i.e. the voltage across the secondary is greater than the input primary voltage. Such a transformer in which the number of turns in the secondary is more than in the primary is called a *step-up* transformer. But if $N_s < N_p$; $\mathcal{E}_s < \mathcal{E}_p$. Such a transformer is called a *step-down* transformer. The former are used in TV, high voltage power supplies and the latter in radio transmitter sets, battery eliminators, etc.

Usually, there are a number of energy losses in actual transformers. These are: (i) joule heating (I^2R) losses in the primary and secondary coils due to their resistance (generally, these losses are minimized by using wires of large diameters so that resistance is low); and (ii) the losses in the iron core which include the heating of the core due to eddy currents and power loss due to hysteresis. The eddy currents can be minimized by using laminated iron.

The most useful application of transformers is in the transmission of electrical power from a power station (where it is generated) to far away cities and towns (where it is used). Very long cables are needed to carry electricity from the power station to the users. A sizeable fraction of power is lost as Joule heat in the cables which are hundreds of kilometres long.

If I is the current in the cable and R its resistance, then the power lost in the cable is I^2R. Thus the power loss can be reduced by reducing I or by reducing R. We have no control on R but we can reduce I. The power supplied by the generator is given by $P = VI$, where V is the voltage across its terminals. Since $I = P/V$, for a given amount of power P, the power loss will be less if I is less, i.e. if V is increased. Hence, to minimize power loss during transmission, the voltage at the generating station is increased to a very high value (about 33,000 V) by using a step-up transformer and delivered to sub-stations in cities. At the sub-station, the voltage is reduced to a safe value ($\sim$ 220 V) by using a step-down transformer. The power is then delivered to individual users. While travelling in a train, you must have seen the overhead transmission cables; they have voltages of the order of a hundred thousand volts.

EXAMPLE 7.13 The resistance of the armature of a generator is 0.2 Ω. It yields an emf of 220 V in an open circuit and a potential difference of 210 V at full load. Calculate:

(a) the current at full load; and
(b) the power delivered to the external circuit.

Solution:

Resistance of armature = 0.2 Ω
Potential difference in open circuit = 220 V
Potential difference at full load = 210 V

(a) Current in the circuit (I) $= \dfrac{220 - 210}{0.2} = 50$ A

(b) Power delivered $= 210 \times 50 = 10.5$ kW

EXAMPLE 7.14 A step-down transformer is used to reduce the main supply of 220 V to 10 V. If the primary draws 5 A and the secondary 100 A, calculate the efficiency of the transformer.

Solution:

Primary voltage = 220 V
Primary current = 5 A

Secondary voltage = 10 V

Secondary current = 100 A

Power input = 220 × 5 = 1100 W

Power output = 100 × 10 = 1000 W

$$\text{Efficiency} = \frac{\text{power output}}{\text{power input}} = \frac{1000}{1100} = 0.91 \text{ or } 91\%$$

EXAMPLE 7.15 (a) The primary of a transformer has 400 turns while the secondary has 2000 turns. If the power output from the secondary at 1100 V is 12.1 kW, calculate the primary voltage.

(b) If the resistance of the primary is 0.2 Ω and that of the secondary is 2.0 Ω, and the efficiency of the transformer is 90%, calculate

(i) the heat loss in the primary coil; and
(ii) the heat loss in the secondary coil.

Solution:

(a) $N_p = 400 \qquad N_s = 2000$

$\mathcal{E}_s = 1100$ V

$$\therefore \quad \mathcal{E}_p = 1100 \times \frac{400}{2000} = 220 \text{ V}$$

∴ Primary voltage = 220 V

(b) Resistance of primary = 0.2 Ω

Resistance of secondary = 2.0 Ω

Power output = 12.1 kW = 12100 W

Efficiency = 90%

$\mathcal{E}_s = 1100$ V

$\mathcal{E}_p = 220$ V

$$\text{Current in the secondary} = \frac{12100}{1100} = 11 \text{ A}$$

Since the efficiency is 90%, the input power is

$$12.1 \times \frac{100}{90} = 13.44 \text{ k}$$

Current in the primary

$$= \frac{13.44 \times 10^3}{220}$$

$$= 61.1 \text{ A}$$

(i) Power loss in the primary $= (61.1)^2 \times 0.2 = 747$ W
(ii) Power loss in the secondary $= (11)^2 \times 2.0 = 242$ W

EXAMPLE 7.16 A power transformer is used to step up an alternating emf of 220 V to 4.4 kV to transmit 6.6 kW of power. If the primary has 1000 turns, find: (a) the number of turns in the secondary, and (b) the current rating of the secondary. Assume a 100% efficiency for the transformer.

Solution: $\mathcal{E}_p = 220$ V $\quad \mathcal{E}_p = 4.4$ kV $= 4.4 \times 10^3$ V

$N_p = 1000 \quad$ Power $(P) = 6.6$ kW $= 6.6 \times 10^3$ W

(a) $$N_s = \left(\frac{\mathcal{E}_s}{\mathcal{E}_p}\right) N_p = \frac{4.4 \times 10^3 \times 1000}{220} = 20{,}000$$

(b) Let I_s be the current in secondary and I_p that in the primary. For a transformer of a 100% efficiency (see supplement S.2 at the end of this chapter)

$$I_s \mathcal{E}_s = I_p \mathcal{E}_p = P = 6.6 \times 10^3 \text{ W}$$

$$\therefore \qquad I_s = \frac{P}{\mathcal{E}_s} = \frac{6.6 \times 10^3}{4.4 \times 10^3} = 1.5 \text{ A}$$

Thus the current rating of the secondary is 1.5 A.

SUPPLEMENTS

S.1 ENERGY STORED IN AN INDUCTOR

Just as energy is stored in the electrostatic field of a charged capacitor, energy is stored in the magnetic field of a current carrying inductor. When the current in the coil is first switched on, the induced emf opposes the rise of current, the current flowing against the induced emf, therefore, does work against it. When the current becomes steady, there is no induced emf and no more work is done against it. The total work done in bringing the current to its final steady value is stored in the magnetic field of the coil as magnetic potential energy.

To calculate the energy stored in the inductor, suppose that the current through it is rising at a rate dI/dt. Then if L is the self-inductance, the induced emf is (numerically) given by

$$|\mathcal{E}| = L\frac{dI}{dt}$$

If, at that time, the value of current is I, then the rate at which work is being done against the induced emf is

$$P = \mathcal{E} I = LI\frac{\mathrm{d}I}{\mathrm{d}t}$$

The total work done in bringing the current from zero to a steady value I_0 is, therefore,

$$W = \int_0^{I_0} P\,\mathrm{d}t = \int_0^{I_0} LI\,\frac{\mathrm{d}I}{\mathrm{d}t}\cdot\mathrm{d}t = L\int_0^{I_0} I\,\mathrm{d}t = \frac{1}{2}LI_0^2 \qquad (7.22)$$

This energy is stored in the magnetic field of the coil.

S.2 CURRENTS AND VOLTAGES IN A TRANSFORMER

Consider an 'ideal' transformer which is defined as one in which the magnetic field lines are confined entirely within the core, with no field outside the core. If N_p and N_s are the number of turns in the primary and the secondary and if Φ is the flux across a single turn, then the fluxes in the coils are

$$\Phi_p = N_p\Phi \qquad \Phi_s = N_s\Phi$$

From Faraday's law, the voltages across the coils are

$$V_p = -\frac{\mathrm{d}\Phi_p}{\mathrm{d}t} = -N_p\frac{\mathrm{d}\Phi}{\mathrm{d}t}$$

and

$$V_s = -\frac{\mathrm{d}\Phi_s}{\mathrm{d}t} = -N_s\frac{\mathrm{d}\Phi}{\mathrm{d}t}$$

Hence

$$\frac{V_p}{V_s} = \frac{N_p}{N_s} \quad \text{or} \quad V_s = nV_p \qquad (7.23)$$

where $n = N_p/N_s$ is the turns ratio.

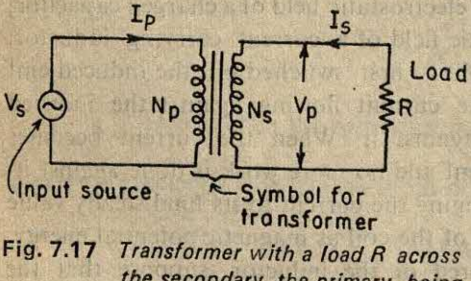


Fig. 7.17 *Transformer with a load R across the secondary, the primary being fed by an ac source*

The primary is connected to an ac voltage source $V_p = V_0 \sin \omega t$ and the secondary has a load R across it (Fig. 7.17).

Then

$$V_s = nV_0 \sin \omega t \qquad (7.24)$$

and the current in the secondary is

$$I_s = \frac{V_s}{R} = \frac{nV_0}{R} \cdot \sin \omega t \tag{7.25}$$

As a result of the current I_s, an additional flux is produced in the core of the transformer. By Lenz's law, this flux opposes the original flux due to I_p. So the flux in the core will be

$$\Phi = (N_p I_p - N_s I_s)\Phi_0$$

where Φ_0 denotes the flux in the core *per unit current* due to a *single turn*. We have assumed that the magnetic field of each coil is confined to the core. Hence

$$\Phi_p = N_p \Phi = N_p \Phi_0 (N_p I_p - N_s I_s)$$

But $$V_p = V_0 \sin \omega t = -\frac{d\Phi_p}{dt}$$

$$= -N_p \Phi_0 \left(N_p \frac{dI_p}{dt} - N_s \frac{dI_s}{dt} \right)$$

or $$N_p \frac{dI_p}{dt} - N_s \frac{dI_s}{dt} = -\frac{V_0}{N_p \Phi_0} \cdot \sin \omega t$$

Integrating we get

$$N_p I_p - N_s I_s = + \frac{V_0 \cos \omega t}{N_p \omega \Phi_0}$$

Using Eq. (7.25) we have

$$I_p = \frac{V_0 \cos \omega t}{N_p \omega \Phi_0} + \frac{n^2 V_0 \sin \omega t}{R} \tag{7.26}$$

Equations (7.24), (7.25) and (7.26) give the voltage across the secondary and currents in the two coils; the voltage across the primary being $V_p = V_0 \sin \omega t$.

The total power supplied, i.e. power input by the source during one complete cycle from $t = 0$ to $t = T$ is (where period $T = 2\pi/\omega$)

$$P_{in} = \int_0^T V_p I_p \, dt = \int_0^T (V_0 \sin \omega t) \left(\frac{V_0 \cos \omega t}{N_p \omega \Phi_0} + \frac{n^2 V_0 \sin \omega t}{R} \right) dt$$

$$= \frac{V_0^2}{N_p \omega \Phi_0} \int_0^T (\sin \omega t \cdot \cos \omega t) \, dt + \frac{n^2 V_0^2}{R} \int_0^T (\sin^2 \omega t) \, dt \tag{7.27}$$

Now $$\int_0^T (\sin \omega t \cdot \cos \omega t) \, dt = \frac{1}{2} \int_0^T (\sin 2\omega t) \, dt$$

$$= \frac{1}{2}\int_0^T \sin\left(\frac{4\pi t}{T}\right)\mathrm{d}t$$

$$= -\frac{1}{2}\,\frac{T}{\pi}\cdot\cos\,\frac{4\pi t}{T}\bigg|_0^T = 0$$

and

$$\int_0^T (\sin^2 \omega t)\,\mathrm{d}t = \int_0^T \frac{1}{2}(1 - (\cos 2\omega t)\,\mathrm{d}t$$

$$= \frac{1}{2}\int_0^T \mathrm{d}t - \frac{1}{2}\int_0^T (\cos 2\omega t)\,\mathrm{d}t$$

$$= \frac{T}{2} - 0 = \frac{T}{2} = \frac{\pi}{\omega}$$

Using these values of integrals in Eq. (7.27) we get

$$P_{\text{in}} = \frac{\pi n^2 V_0^2}{R\omega} \tag{7.28}$$

The power output (i.e. power dissipated at the load) is

$$P_{\text{out}} = \int_0^T V_s I_s\,\mathrm{d}t = \frac{n^2 V_0^2}{R}\int_0^T (\sin^2 \omega t)\,\mathrm{d}t$$

$$= \frac{\pi n^2 V_0^2}{R\omega} \tag{7.29}$$

Notice that $P_{\text{in}} = P_{\text{out}}$ as expected, as the transformer has been assumed to be ideal. In an actual transformer, the power output is less than power input due to eddy currents in the core, also hysteresisof the core and, of course, Joule heating. To reduce loss, the core is laminated (thin layers separated by insulating barriers). This increases the resistance of the core and hence decreases the eddy currents.

SUMMARY

Whenever there is a change in the magnetic flux linked with a circuit, an induced emf and induced current are set up. This phenomenon is known as electromagnetic induction. Most of the commercial sources of electric energy use this phenomenon. Faraday's law of electromagnetic induction states that the induced emf is numerically proportional to the rate of change of flux. Lenz's law of electromagnetic induction gives the direction of the induced current and it states that the induced emf is always in such a direction so as to oppose the very cause that gives rise to it. Combining these two laws, the induced emf can be written as

$$\mathcal{E} = -\frac{d\Phi}{dt}$$

The induced emf is set up even if the circuit is not closed. The emf induced in a straight wire of length l moving with a velocity $\mathbf{v}$ in a uniform magnetic field $\mathbf{B}$ is given by

$$\mathcal{E} = Blv \sin \theta$$

where θ is the angle between $\mathbf{v}$ and $\mathbf{B}$.

Whenever there is an induced current in a circuit, the charge flowing at any point in the circuit depends on the change in flux only and is independent of time.

If a coil is rotated with angular velocity ω in a uniform magnetic field, the instantaneous emf developed at the ends of the coil is

$$\mathcal{E} = \mathcal{E}_0 \sin \omega t$$

where $\mathcal{E}_0$ is the maximum emf and is equal to $NAB\omega$. This is the basis of an ac generator.

Generators transform mechanical energy into electric energy. A simple generator consists of a rectangular coil of wire rotating in a uniform magnetic field. The induced emf in such a generator is alternating. Eddy or Foucault currents are the currents induced in large conducting bodies when they are linked with changing magnetic fluxes. A transformer is a device used for increasing or decreasing the ac voltages. It consists of two coils, called primary and secondary, wound over an iron core.

Electric energy is transmitted from a power house to long distances at high voltage and low current in order to reduce the I^2R heating losses in the transmission line considerably.

EXERCISES

A. Short-Answer Questions

1. State the SI unit of magnetic flux.
2. What is the SI unit of the rate of change of flux with time?
3. What are eddy currents?
4. Define self-inductance and state its SI unit.
5. Define a henry.

B. Long-Answer Questions

1. (a) What do you understand by the phenomenon of electromagnetic induction?
   (b) What conditions must be satisfied to observe this phenomenon?

2. State the laws of electromagnetic induction and write them mathematically.
3. (a) Explain how Lenz's law follows from the law of conservation of energy.
   (b) Give two practical examples where Lenz's law is applied.
4. Why do the lights of the house become dim as soon as the refrigerator is switched on?
5. (a) Explain briefly from the free electron theory of metals how an induced emf is produced in a moving wire when placed in a magnetic field.
   (b) What is the value of induced emf if the wire moves in the direction of the magnetic field.
6. (a) Describe briefly the principle of an ac generator. Can such a current be measured with a moving coil galvanometer?
   (b) Explain the statement, "every electric motor is necessarily a generator also."
   (c) Why is it easier to turn the shaft of a dynamo or ac generator in an open circuit than in a closed circuit?
   (d) Explain why an electric motor draws more current when it starts running than when it runs continuously.
7. (a) What are eddy currents and how are they minimized?
   (b) Give one example where the eddy currents are: (i) desirable, and (ii) undesirable.
8. (a) Explain the principle of a transformer. Can a transformer work when connected to a battery?
   (b) What are step-up and step-down transformers? Give two examples of each.
   (c) Mention the power losses in a transformer.
9. Explain with the help of an example why ac is used for the long-distance transmission of electric power.
10. What do you understand by self-induction?
11. (a) Define self-inductance and give its SI unit.
    (b) What is the limitation of the definition?
    (c) What is an ideal inductor? Can it be realized in practice?
    (d) Calculate the self-inductance of an ideal solenoid.
    (e) Does the coil of a moving coil galvanometer have any inductance? If it does, why does it not affect its working?
    (f) Show that the equation

$$\mathcal{E} = -L\frac{\mathrm{d}I}{\mathrm{d}t}$$

    is dimensionally valid.
12. (a) Define mutual inductance.
    (b) What are the factors on which the self- and mutual inductance depend?

C. Multiple-Choice Questions

Choose the correct answer/answers from the given alternatives.

1. The magnitude of the induced emf produced in a coil when a magnet is inserted into it does not depend upon the

(a) number of turns in the coil (b) resistance of the coil
(c) magnetic moment of the magnet (d) speed of approach of the magnet

2. A coil of wire is held with its plane horizontal to the earth's surface and a small bar magnet dropped vertically down through it. The magnet will fall with a

   (a) constant acceleration equal to g
   (b) constant acceleration greater than g
   (c) constant acceleration less than g
   (d) non-uniform acceleration less than g.

3. Two identical coils of insulated wire are hung freely, facing each other. If they are fed exactly in the same way with an alternating current, they will

   (a) attract each other
   (b) repel each other
   (c) rotate in the clockwise direction
   (d) rotate in the anti-clock wise direction.

4. Transformers are used in

   (a) dc circuits alone (b) ac circuits alone
   (c) both ac and dc circuits (d) integrated circuits.

5. A transformer has 200 windings in the primary and 400 windings in the secondary. The primary is connected to an ac supply of 110 V and a current of 10 A flows in it. The voltage across the secondary and the current in it respectively, are

   (a) 55 V, 20 A (b) 440 V, 5 A
   (c) 220 V, 10 A (d) 220 V, 5 A

6. The magnitude of the emf across the secondary of a transformer does not depend upon

   (a) the magnitude of the emf applied across the primary
   (b) the number of turns in the primary
   (c) the number of turns in the secondary
   (d) the resistances of the primary and the secondary.

7. Electrical energy generated at a power house is delivered to distant places over long transmission cables at a very high ac voltage of about 330,000 volts. The reason for this is that

   (a) at high voltages energy is delivered much faster than at low voltages
   (b) there is less wastage of energy at high voltages
   (c) the high voltage prevents theft of costly transmission cables
   (d) it is much easier to generate large amounts of energy at high voltages

8. Flux Φ (in weber) in a closed circuit of resistance 10 Ω varies with time t (in seconds) according to the equation

$$\Phi = 6t^2 - 5t + 1$$

The magnitude of the induced current in the circuit at $t = 0.25$ s is

(a) 0.2 A (b) 0.6 A
(c) 0.8 A (d) 1.2 A

9. The emf induced in a 1 millihenry inductor in which the current changes from 5 A to 3 A in 10^{-3} is

   (a) 2×10^{-6} V (b) 8×10^{-6} V
   (c) 2 V (d) 8 V

10. A circuit has a self-inductance of 1 henry and carries a current of 2 A. To prevent sparking when the circuit is broken, a capacitor which can withstand 400 V is used. The least capacitance of the capacitor connected across the switch must be equal to

    (a) 12.5 μF (b) 25 μF
    (c) 50 μF (d) 100 μF

11. An aeroplane is moving north horizontally, with a speed of 200 ms^{-1}, at a place where the vertical component of the earth's magnetic field is 0.5×10^{-4} T. What is the induced emf set up between the tips of the wings if they are 10 m apart?

    (a) 0.01 V (b) 0.1 V
    (c) 1 V (d) 10 V

12. A step-down transformer is employed to reduce the main supply of 220 V to 11 V. The primary draws 5 A of current and the secondary 90 A. What is the efficiency of the transformer?

    (a) 20% (b) 40%
    (c) 70% (d) 90%

D. Numerical Problems

1. Flux Φ (in weber) in a closed circuit of resistance 10 Ω varies with time t (in seconds) according to the equation

$$\Phi = 6t^2 - 5t + 4$$

Calculate: (i) the induced emf, and (ii) the induced current when $t = 0.25$ s.

2. A rectangular coil of 50 turns and area 0.05 m^2 rotates in a magnetic field of 0.1 T with an angular velocity of 100 rad s^{-1}. If the coil starts from position of zero emf, calculate

   (a) the average emf for 90° rotation,
   (b) the average emf for the next half revolution, i.e. 90° to 270°, and
   (c) instantaneous emf when it has rotated through 30°

3. A circular wire of diameter 0.20 m and resistance 1.0 Ω is placed in a uniform magnetic field. If a charge of 2.0 C passes through any point, calculate the change in the flux linked with the wire.

4. (a) A wire of length 1 m moves with a constant velocity of 5 ms^{-1} in a uniform magnetic field 0.2 Wb m^{-2} acting perpendicular to the wire. Calculate the induced emf in the wire if its direction of motion makes an angle of 60° with the magnetic field.

   (b) A circus trapeze consists of a metal bar 0.80 m long suspended by two ropes each of length 60 m. Calculate the maximum induced emf developed at the ends of the bar if the bar is drawn back until it is at the same level as the supporting hooks and then released. Given that the vertical component of earth's field is 0.5×10^{-4} Wb m^{-2}.

(c) If the distance between the rail tracks is 1.0 m and vertical component of the earth's magnetic field is 0.6×10^{-4} T, (i) calculate the induced emf between the rails when a train is travelling with a velocity of 72 km hr^{-1} on the track. (ii) What is the retarding force on the train if the resistance between the rail tracks is 1.00 Ω?

(d) A copper disc of radius 7 cm rotates about an axis perpendicular to its plane with a frequency of 1800 rpm in a magnetic field of induction 0.4 T acting along the axis of the disc. What is the reading in a millivoltmeter connected to two brushes in contact with the edge and the axis of the disc?

5. (a) The armature of an ac generator has a coil of 1000 turns each of area 0.2 m^2. The coil is rotated at 1200 rpm in an uniform magnetic field of 0.01 T.

   (i) What is the maximum emf induced?
   (ii) Write the induced emf as a function of time.
   (iii) What is the instantaneous emf one second after the coil passes a position of zero emf?

   (b) The emf from an ac generator is given by

$$\mathcal{E} = 300 \sin\left(314t + \frac{\pi}{4}\right)$$

   What is: (i) the amplitude, (ii) the frequency of rotation, and (iii) the phase angle?

   (c) In a dynamo, a coil of area 0.2 m^2 and 100 turns rotates in a magnetic field of 0.1 Wb m^{-2} with an angular velocity of 100 rad s^{-1}. The output terminals are connected through a resistance of 100 Ω. Calculate the torque τ, as a function of time t, required to keep the coil rotating.

6. (a) A ballistic galvanometer has a resistance of 100 Ω and it gives a full-scale deflection for a charge of 1 μC. A circular coil of 200 turns and 50Ω resistance is to be constructed to study magnetic fields up to 1 T. Calculate the maximum allowable radius of the coil.

   (b) A closely wound search coil of 30 turns (area 1 cm^2 and resistance 4 Ω is connected through a wire of negligible resistance to a ballistic galvanometer of 20 Ω resistance. If the coil is suddenly removed from the region of magnetic field of 0.5 T to a region of zero field, calculate the charge flowing in the galvanometer when the plane of coil, while in the magnetic field:

   (i) is perpendicular to field direction
   (ii) makes an angle of 30° with the field direction.

7. (a) A transformer of 100% efficiency has 200 turns in the primary and 40,000 turns in the secondary. It is connected to a 220 V main supply and the secondary feeds to a 100 kΩ resistance. Calculate

   (i) the output potential difference per turn;
   (ii) the total output potential difference; and
   (iii) the power delivered to load.

(b) A step-down transformer having a power output of 10 kW and efficiency 90% reduces the voltage from 11 kV to 220 V. Calculate:

(i) the number of turns in the primary if the secondary has 100 turns; and
(ii) the current in the primary.

(c) An ac generator in a power house generates electricity at 440 V and 200 A. It is stepped up to 20 times with the help of a transformer and transmitted to distant places by a transmission line. If the resistance of the transmission line is 1 Ω per km, calculate:

(i) the percentage power loss if there was no step-up transformer at a place 1 km away from the power house; and
(ii) the percentage power loss in the presence of the transformer for the same place as in (i).

8. (a) A solenoid of radius 3 cm has 600 turns per metre. What is its self-inductance?

(b) If this inductance is wound on an iron piece, will the value of inductance change?

9. (a) An ideal solenoid of 1000 turns has an inductance of 1 H.

(i) Calculate the energy stored in it when a current of 4 A is passed through it.
(ii) If another coil of 50 turns is closely wound around and near the centre of this solenoid, calculate the mutual inductance between the coil and the solenoid.

(b) (i) What is the self-inductance of a toroid which has 200 turns, the inner and outer radii being 7.5 and 9.0 cm respectively.
(ii) If another coil of 50 turns is wrapped closely around this toroid, what will be the mutual inductance between the two coils?

10. An aircraft with a wing span of 40 m flies with a speed of 1000 km h^{-1} in the eastward direction at a constant altitude in the northern hemisphere where the horizontal component of earth's field is 1.6×10^{-5} T and the angle of dip is 71.6°. Find the emf that develops between the tips of the wings of the aircraft.

11. Two coils have a mutual inductance of 5×10^{-3} H. The current in one coil changes according to the relation $I = I_0 \sin \omega t$, where $I_0 = 10$ A, $\omega = 2\pi/T$ and $T = 0.02$ s. Find: (a) the time dependence of the induced emf in the other coil, and (b) the maximum value of the induced emf.

8

ALTERNATING CURRENT CIRCUITS

8.1 INTRODUCTION

So far we have considered only dc circuits, i.e. circuits in which the primary source is a battery. The basic circuit element in a dc circuit is a resistor (R) which controls the current and voltage in the circuit and the relationship between them is $V = IR$ (Ohm's law). But dc circuits are not the only ones we use. In fact, most of the electrical power generated and used in the world is in the ac form in which the voltage and current vary with time. In ac circuits, in addition to resistors (R), two new circuit elements, namely inductors (L) and capacitors (C) become relevant. The currents and voltages in ac circuits are controlled by three circuit elements L, C and R.

If we have a circuit consisting of a pure inductance L in which the current is changing with time at a rate $\mathrm{d}I/\mathrm{d}t$, then the voltage across the inductor is numerically given by (see Chapter 7)

$$V = L\,\frac{\mathrm{d}I}{\mathrm{d}t} \tag{8.1a}$$

This is the voltage-current relationship for an inductor, similar to the relation $V = RI$ for a resistor.

If we have a circuit consisting of a capacitance C across which the voltage is changing at a rate $\mathrm{d}V/\mathrm{d}t$, then the current through the capacitor is given by

$$I = C\,\frac{\mathrm{d}V}{\mathrm{d}t} \tag{8.1b}$$

which follows from the relations $q = CV$ and $I = \mathrm{d}q/\mathrm{d}t$. This is the currents-voltage relation for the capacitors. Notice from Eqs (8.1a and b) that the current-voltage relationship for inductors and capacitors looks very different from the Ohm's law relationship $V = IR$ for resistors. It is clear that, if I and V do not change with time (as in dc circuits), an inductor or a capacitor does not affect the voltage or current. What effect an inductor or capacitor has in a circuit in which voltage and current are time-dependent is the subject matter of this chapter.

We will consider two kinds of time-varying voltages. First we will discuss what happens when a circuit is switched on or off. We will see (refer to Sec. 8.2) that when an inductor or a capacitor is present in the circuit (in addition to a resistor) the voltage (and hence current) does not rise or fall immediately but does so gradually with a characteristic time called the time constant of the circuit. The second kind of time-varying voltage we will discuss is the one which varies sinusoidally with time, i.e $V(t) = V_0 \sin \omega t$ or $V_0 \cos \omega t$. We will see how inductors and capacitors affect such circuits. In particular, we will show that they behave somewhat like resistors. To do this, we will use the method of complex numbers. Section 8.3 describes a few properties of complex numbers we will need to know. Section 8.4 then discusses ac circuits.

8.2 GROWTH AND DECAY OF CURRENT IN AN INDUCTOR AND A CAPACITOR

Growth of Current in an Inductor

Consider an ideal inductor L and a resistor R connected to a battery of emf $\mathcal{E}$ and negligible internal resistance through a two-way key K (Fig. 8.1). By connecting the terminals 1 and 2 the current I flows in the inductor. The current in the circuit is zero as soon as the terminals 1 and 2 are connected. If there is no inductor, the current in the circuit would rise rapidly to a steady value $\mathcal{E}/R$. Due to the presence of the inductor an induced emf $\mathcal{E}_L = -L\dfrac{dI}{dt}$ appears in the circuit. By Lenz's law, this emf opposes the rise of current in the circuit. Since $\mathcal{E}_L$ is opposite to the emf $\mathcal{E}$ of the battery, the net potential difference that appears across the resistor is $\mathcal{E} - L\dfrac{dI}{dt}$ and by Ohm's law this must be equal Io IR. Hence,

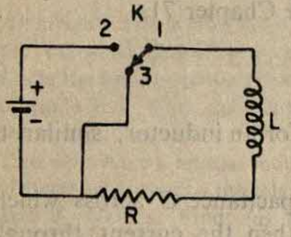


Fig. 8.1 *L and R connected to a battery*

$$\mathcal{E} - L\frac{dI}{dt} = IR$$

or
$$\mathcal{E} = IR + L\frac{\mathrm{d}I}{\mathrm{d}t} \tag{8.1}$$

This is a differential equation in I which can be solved by direct integration and using the fact that $I = 0$ at $t = 0$. The solution is

$$I = \frac{\mathcal{E}}{R}(1 - e^{-Rt/L}) \tag{8.2}$$

From Eq. (8.2), we have on differentiation:

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \frac{\mathcal{E}}{L}e^{-Rt/L} \tag{8.3}$$

Using Eqs (8.2) and (8.3), one can verify that Eq. (8.1) is satisfied. From Eqs (8.2) and (8.3) one gets the expression for the potential difference V_R across the resistance and the potential difference V_L across the inductor as,

$$V_R = IR = \mathcal{E}(1 - e^{-Rt/L})$$

$$V_L = \left| -L\frac{\mathrm{d}I}{\mathrm{d}t}\right| = \mathcal{E}\, e^{-Rt/L}$$

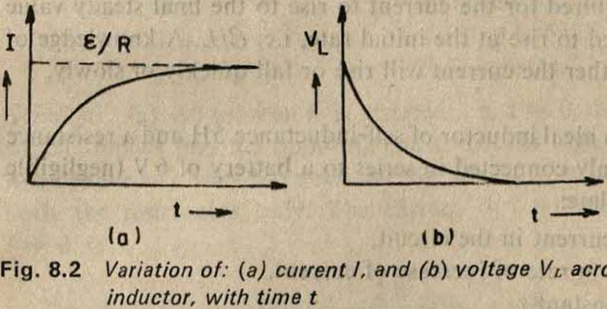


Fig. 8.2 *Variation of: (a) current I, and (b) voltage V_L across inductor, with time t*

The above equations show how I and V_L vary with time t. Figure 8.2(a) shows the variation of I and Fig. 8.2(b) shows the variation of V_L. From the figure it is clear that the current I does not rise to a steady value instantaneously as in a resistance; rather it rises asymptotically to the steady value of current $I_0 = \mathcal{E}/R$. The potential difference V_L is maximum at $t = 0$ and falls off exponentially to a zero value. Physically the whole process can be understood as follows. As the current starts rising, the opposing emf ($L\,\mathrm{d}I/\mathrm{d}t$) acts. As time passes the current in the circuit approaches the steady value $\mathcal{E}/R$ and the rate of rise of current ($\mathrm{d}I/\mathrm{d}t$) falls to zero. So as the current approaches a steady value, the counter emf ($L\,\mathrm{d}I/\mathrm{d}t$) due to the inductor approaches zero as shown in Fig. 8.2(b). Further, it is clear from Eq. (8.2) that whether the growth of current is fast or slow depends upon the values of L and R. If L is low, the current rises more rapidly and soon approaches the steady value. If L is high, the current rises very slowly. Also, it can be seen easily from the units of L and R, that the quantity L/R has the dimensions of time. If we write $\tau = L/R$, τ is called the *time constant* of the LR circuit. The significance of L/R can be understood as follows.

When $t = \tau$, we have

$$\frac{Rt}{L} = 1$$

So from Eq. (8.2)

$$I = \frac{\mathcal{E}}{R}(1 - e^{-1})$$

$$\simeq 0.63 \frac{\mathcal{E}}{R} \qquad (\because \; e = 2.718)$$

or

$$I = 0.63 I_0 \qquad (\because \; \mathcal{E}/R = I_0)$$

τ, therefore, can be defined as the time required for the current to rise to 0.63 of the final steady value $I_0 \cdot \tau$ may also be defined as the time that would be required for the current to rise to the final steady value if it had continued to rise at the initial rate, i.e. $\mathcal{E}/L$. A knowledge of τ determines whether the current will rise or fall quickly or slowly.

EXAMPLE 8.1 An ideal inductor of self-inductance 5H and a resistance of 100 Ω is suddenly connected in series to a battery of 6 V (negligible resistance). Calculate:

(a) the steady current in the circuit,
(b) the maximum rate of increase of current,
(c) the time constant τ
(d) the value of current after $t = 0.5\tau$, and
(e) the potential drop across the inductor at $t=0.5\tau$

Solution:

$$L = 5 \text{ H}$$
$$R = 100 \; \Omega$$
$$\mathcal{E} = 6 \text{ V}$$

(a) Steady current $= \dfrac{\mathcal{E}}{R} = 0.06$ A

(b) The rate of increase of current is maximum for $t = 0$ and it is

$$\frac{\mathcal{E}}{L} = \frac{6}{5}$$
$$= 1.2 \text{ As}^{-1}$$

(c) $\tau = L/R = 0.05$ s

(d) When $t = 0.5\tau$,

$$I = 0.06(1 - e^{-0.5})$$

$$= 0.06(1 - 0.606) = 0.024 \text{ A}$$

(e) Potential drop $= \mathcal{E} - IR = 6 - 100 \times 0.024 = 3.6$ V

EXAMPLE 8.2 Two resistances of 10 and 20 Ω and an ideal inductor of inductance 5 H are connected to a battery of 2 V through a key K as shown in Fig. 8.3. If at $t = 0$, 'K, is inserted calculate:

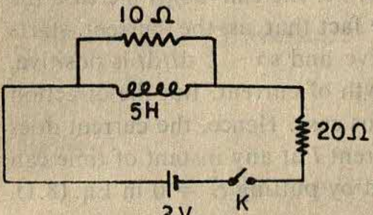


Fig. 8.3

(a) the initial current through the battery,

(b) the initial potential drop across the inductor,

(c) the final current through 10 Ω resistor, and

(d) the final current through 20 Ω resistor.

Solution: (a) As soon as K is inserted, i.e. $t = 0$, dI/dt is maximum, which implies that the opposing emf $L\, dI/dt$ is high and the inductor will behave as a very large resistor. So the current will flow through both the resistances only. The current through the battery $I(0)$ at $t = 0$, is

$$I(0) = \frac{2.0}{10 + 20} = \frac{1}{15} \text{ A}$$

(b) Since the 10 Ω resistor and inductor are in parallel, the potential drop across the inductor is the same as that across the 10 Ω resistor. Hence, the initial potential drop is

$$I(0) \times 10 = \frac{10}{15} \text{ V}$$

(c) When the current has attained a constant value, the opposing emf across the inductor is zero. The inductor would behave as a short and the whole current will pass through it. The final current through the 10 Ω resistor is, therefore, zero.

(d) Since the inductor behaves as a short or zero resistance, the total resistance of the circuit is also 20 Ω (for 10 Ω is in parallel with the inductor). The final current in the circuit is, therefore,

$$2.0/20 = 0.1 \text{ A}.$$

Decay of Current Now we shall study how the current behaves when the battery is suddenly short-circuited (by connecting terminals 1 and 3 of key K as shown in Fig. 8.3) after the current has attained a steady value. As soon as the battery is short-circuited, the LR circuit is again closed and the current would decrease from its steady value to zero, unlike the case of a resistance in which the current falls off instantaneously. Here, as soon as the current starts decreasing, an induced emf $-L\,\mathrm{d}I/\mathrm{d}t$ across the inductor appears and this opposes the decrease of current (by Lenz's law), i.e. the emf helps increase the current. This can easily be seen by the fact that as the current starts decreasing with time, $\mathrm{d}I/\mathrm{d}t$ is negative and so $-L\,\mathrm{d}I/\mathrm{d}t$ is positive, unlike in the previous case of the growth of current. But the direction of current is the same as in the previous case. Hence, the current does not fall off instantaneously. The current I at any instant of time can now be given by the equation obtained by putting $\mathcal{E} = 0$ in Eq. (8.1)

$$IR + L\frac{\mathrm{d}I}{\mathrm{d}t} = 0 \tag{8.4}$$

The solution of Eq. (8.4) is

$$I = \frac{\mathcal{E}}{R}e^{-t/\tau} \tag{8.5}$$

where τ time constant of the circuit. From Eq. (8.5) it is clear that the current falls off exponentially as shown in Fig. 8.4. The rate of this fall depends on the values of L and R. If L is large, the time constant is also large, so that the current would decrease very slowly. It is because of this slow decrease of current that there is a spark across the terminals of a switch or key when a dc circuit containing a large inductor is just, or partially, opened. This spark can damage the switch contacts and insulation, and can be dangerous to the circuit. Because of this reason, the switches in a large inductance circuit are usually immersed in oil to prevent damage due to sparking.

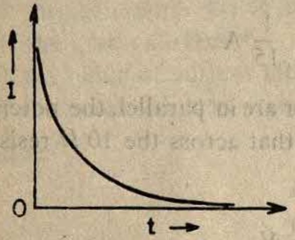


Fig. 8.4 *Decay of current with time*

EXAMPLE 8.3 An inductor of self-inductance 100 mH and resistance 50 Ω is connected to a battery of 2 V through a two-way key. The battery is connected in the circuit. After a long time when the current has attained a steady value the battery is short-circuited. Calculate:

(a) the current after 0.01 s, and

(b) the time required for the current to fall to half its steady value.

Solution: $L = 100 \times 10^{-3}$ H $\quad R = 50\,\Omega \quad \mathcal{E} = 2$ V

$$\text{Steady current} \quad I_0 = \frac{2}{50} = 0.04 \text{ A}$$

$$\text{Time constant} \quad \tau = \frac{L}{R} = 0.002 \text{ s}$$

(a) $t = 0.01$ s. Therefore,

$$\frac{t}{\tau} = \frac{0.01}{0.002} = 5$$

Hence

$$I = I_0 e^{-5} = 0.04 \times e^{-5} = 268\ \mu\text{A}$$

(b) $$I = I_0 e^{-t/\tau}$$

Let at $t = t_1$, $I = I_0/2$. Hence,

$$e^{-t_1/\tau} = \frac{1}{2}$$

or $$e^{t_1/\tau} = 2 \quad \text{or} \quad \frac{t_1}{\tau} = \ln 2 = 0.693$$

$\therefore$ $$t_1 = \tau \times 0.693 = 2 \times 0.693 = 1.386 \text{ millisecond}$$

Growth and Decay of Charge in a Capacitor

Similar to the case of the growth and decay of current in the case of an inductor, one can also study the growth and decay of charge in a capacitor of capacity C when it is is connected to a battery (emf $\mathcal{E}$ and negligible internal resistance) and a resistor R through a two-way key K as shown in Fig. 8.5.

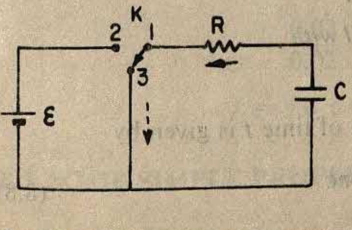


Fig. 8.5 *Capacitor and resistor connected to a battery*

Growth of Charge The battery is introduced in the circuit by connecting terminals 1 and 2 of the key. Initially, i.e. when $t = 0$, the charge on the capacitor plate is zero. As time passes, the charge flows into the capacitor plates and the potential difference q/c (q is the charge at any time t) between the plates rises. The charge on the plates of the capacitor rises till the potential difference between the plates becomes $\mathcal{E}$. The maximum charge collected is $C\mathcal{E}$. At this stage the current in the circuit becomes zero. To see how the charge q, collected on the plates, varies with time t, the equation of charge and current at any instant of time has to be written. If I be the current at any time t, the emf $\mathcal{E}$ of the battery must be equal to the sum of the potential drop IR across the resistor and the potential drop q/C across the capacitor. Therefore,

$$\mathcal{E} = IR + \frac{q}{C} \tag{8.6}$$

If we write $I = \mathrm{d}q/\mathrm{d}t$, Eq. (8.6) is of the same type as in the case of growth of current in an inductive circuit. The charge is given by

$$q = q_0(1 - e^{-t/RC}) \tag{8.7}$$

where $q_0 = \mathcal{E}C$ = maximum charge collected. It is clear from Eq. (8.7) that the charge rises exponentially to a steady value $q_0 = \mathcal{E}C$. The variation of q with t is shown in Fig. 8.6(a). As in the case of the inductive circuit, one can see that the factor RC has the dimension of time and the charge rises to 0.63 of the maximum value in time $t = RC$. Whether the charge rises quickly or slowly depends on the values of R and C. The product RC is known as the time constant of the RC circuit. If RC is very small, the charge rises quickly.

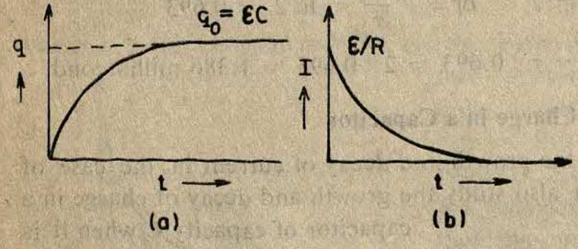


Fig. 8.6 *Variation of charge q and current I with time t*

The behaviour of current as a function of time t is given by

$$I = \frac{\mathrm{d}q}{\mathrm{d}t} = I_0 e^{-t/RC} \tag{8.8}$$

where $I_0 = \mathcal{E}/R$ = maximum current. Therefore, the current decreases exponentially from its maximum value I_0 to zero as shown in Fig. 8.6(b).

Decay of Charge When the charge has attained a steady value $\mathcal{E}C$, the battery is short-circuited by connecting the terminals 1 and 3 of the key K. In such a situation the capacitor starts discharging through the resistor, i.e. the charge on the capacitor starts flowing back through the resistor. The direction of the current is, therefore, reversed as shown by the broken arrow in Fig. 8.5. As in the case of an inductive circuit, the equation governing the variation of charge q with time t is given by [putting $\mathcal{E} = 0$ in Eq. (8.6)]

$$\frac{q}{C} + IR = 0 \tag{8.9}$$

Putting $I = \mathrm{d}q/\mathrm{d}t$ and solving Eq. (8.9) we get

$$q = q_0\,\mathrm{e}^{-t/RC} \tag{8.10}$$

So the charge decays exponentially with time t (Fig. 8.7). Whether the charge decays slowly or quickly depends on the value of the time constant RC.

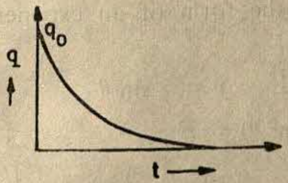


Fig. 8.7 *Decay of charge with time*

EXAMPLE 8.4 A capacitor of 2.0 μF is connected to a battery of 2.0 V through a resistance of 10 kΩ. Calculate:

(a) the initial current in the circuit, and
(b) the current after 0.02 s.

Solution: (a) Initial current in the circuit is

$$I_0 = \frac{\mathcal{E}}{R} = \frac{2}{10\times 10^3} = 0.2\ \mathrm{mA}$$

(b) $$RC = (10\times 10^3)\times(2.0\times 10^{-6}) = 0.02\ \mathrm{s}$$

Now $$t = 0.02\ \mathrm{s}$$

$$\therefore \quad \frac{t}{RC} = \frac{0.02}{0.02} = 1$$

$$\therefore \quad I = I_0 e^{-1} = 0.074\ \mathrm{mA}$$

8.3 SOME SIMPLE PROPERTIES OF COMPLEX NUMBERS

Different mathematical tools have been used to study ac circuits and compute currents and voltages in them. The simplest mathematical method to deal with such circuits is to use the properties of complex numbers and functions. We will first learn the relevant properties in this section and use them in subsequent sections.

If x and y are real numbers then the combination $z = x + iy$ is a complex number. Here $i(= \sqrt{-1})$ is the square root of (-1). Similarly, if α is a real angle, the combination $(\cos\alpha + i\sin\alpha)$ is a complex number. This combination is an exponential function of $i\alpha$, i.e.

$$\exp(i\alpha) = \cos\alpha + i\sin\alpha \tag{8.11}$$

The real and imaginary parts of exp $(i\alpha)$ clearly are

$$R_e[\exp(i\alpha)] = \cos\alpha \tag{8.12}$$

$$I_m[\exp(i\alpha)] = \sin\alpha \tag{8.13}$$

The exponential function exp $(i\alpha)$ is important because any complex number $z = x + iy$ can be written in the form of an exponential function as follows:

Let $\quad x = r\cos\theta \quad$ and $\quad y = r\sin\theta$

Then $\quad x^2 + y^2 = r^2(\cos^2\theta + \sin^2\theta) = r^2$

Therefore $\quad r = (x^2 + y^2)^{1/2} \quad (8.14)$

In terms r and θ, z is given by

$$z = r\cos\theta + ir\sin\theta$$
$$= r(\cos\theta + i\sin\theta)$$

or

$$z = r\exp(i\theta) \tag{8.15}$$

The number r is called the magnitude and θ is the phase of z. The value of θ is given by

$$\tan\theta = \frac{y}{x} \quad \text{or} \quad \theta = \tan^{-1}\left(\frac{y}{x}\right) \tag{8.16}$$

Some important properties of exponential functions are as follows:

1. If α and β are real numbers, then the product

$$\begin{aligned}\exp(i\alpha)\exp(i\beta) &= (\cos\alpha + i\sin\alpha)(\cos\beta + i\sin\beta)\\ &= (\cos\alpha\cos\beta - \sin\alpha\sin\beta) + i(\cos\alpha\sin\beta\\ &\quad + \cos\beta\sin\alpha) \qquad (\because i\times i = -1)\\ &= \cos(\alpha+\beta) + i\sin(\alpha+\beta)\\ &= \exp[i(\alpha+\beta)] \end{aligned} \tag{8.17}$$

i.e. *the product of two (or more) exponential functions is an exponential function whose exponent is the sum of the exponents of the individual functions.*

2. The derivative of an exponential function exp ($i\alpha$) is

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\alpha}[\exp(i\alpha)] &= \frac{\mathrm{d}}{\mathrm{d}\alpha}(\cos\alpha + i\sin\alpha)\\ &= -\sin\alpha + i\cos\alpha\\ &= i(\cos\alpha + i\sin\alpha) \qquad (\because i\times i = -1)\\ &= i\exp(i\alpha) \qquad (8.18)\end{aligned}$$

i.e. *the derivative of the exponential function exp* ($i\alpha$) *is simply i times the function itself.*

3. The integral of an exponential function exp ($i\alpha$) is

$$\begin{aligned}\int \exp(i\alpha)\,\mathrm{d}\alpha &= \int(\cos\alpha + i\sin\alpha)\,\mathrm{d}\alpha\\ &= \int\cos\alpha\,\mathrm{d}\alpha + i\int\sin\alpha\,\mathrm{d}\alpha\\ &= \sin\alpha - i\cos\alpha + \text{constant (of integration)}\\ &= -i(\cos\alpha + i\sin\alpha) + \text{constant}\\ &= -i\exp(i\alpha) + \text{constant} \qquad (8.19)\end{aligned}$$

i.e. *apart from the integration constant, the integral of the exponential exp* ($i\alpha$) *is simply* ($-i$) *times the function itself.* The above mentioned three properties of complex exponential functions considerably simplify the mathematics of ac circuits, as we shall presently see.

8.4 AC CIRCUITS WITH RESISTANCE, INDUCTANCE AND CAPACITANCE

In this section we will compute the current and voltage in a circuit consisting of an ac source and a resistor, an inductor and a capacitor separately or all together. We will assume that the source voltage is given by

$$V = V_0 \cos\omega t \qquad (8.20)$$

where V is the instantaneous voltage at an instant of time t, V_0, the peak value of V (i.e. amplitude or maximum value of V) and ω the angular frequency (in rad s^{-1}) of V. The frequency ν (in hertz) is equal to $\omega/2\pi$ and the period is $T = 1/\nu$.

In any circuit, the currents and voltages are, of course, real quantities. For an ac circuit, instead of writing the source voltage in a real sinusoidal form [Eq. (8.20)], it is mathematically more convenient to consider voltage V as a complex quantity

$$\begin{aligned}V &= V_0\exp(i\omega t) \qquad (8.21)\\ &= V_0(\cos\omega t + i\sin\omega t)\end{aligned}$$

so that the physical voltage $V = V_0 \cos \omega t$ is the real part of the complex voltage V given by Eq. (8.21). The use of a complex voltage will naturally lead to a complex current. The actual (physical) current will be the real part of the complex current.

AC Circuit Containing Only a Resistor

Consider first an ac circuit in which a resistor R is connected across an ac source voltage $V = V_0 \cos \omega t$ (as in the case of domestic supply when mains are connected to a bulb, the frequency of the ac supply in India is $\nu = 50$ Hz so that $\omega = 2\pi\nu = 100\pi$ rad s^{-1}). This is shown in Fig. 8.8(a). To study such a circuit, the use of a complex function is not necessary.

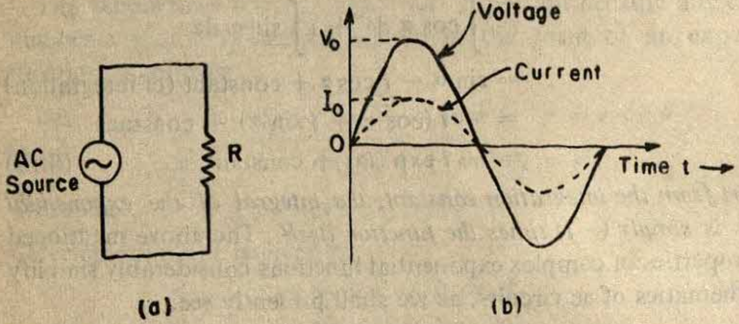


Fig. 8.8 *(a) Ac circuit consisting of a resistor R, and (b) variation of voltage V and current I with time t*

Using Ohm's law, the instantaneous current I is given by

$$I = \frac{V}{R} = \frac{V_0}{R} \cos \omega t = I_0 \cos \omega t \tag{8.22}$$

where $I_0 = V_0/R$ is the peak value of the current. Notice that the ratio

$$\frac{V}{I} = \frac{V_0 \cos \omega t}{I_0 \cos \omega t} = R \tag{8.23}$$

is independent of t, i.e. both V and I vary sinusoidally with time in exactly the same manner; both reach the maximum and zero value at the same time, as shown in Fig. 8.8(b). In other words, *the voltage and current are always in the same phase.*

AC Circuit Containing Only an Inductor

Consider now an ac circuit consisting of a pure (ideal) inductance L [Fig. 8.9(a)]. The voltage applied across the inductor is $V = V_0 \cos \omega t = V_0$ Re $[\exp (i\omega t)]$. Since the voltage is changing with time, the

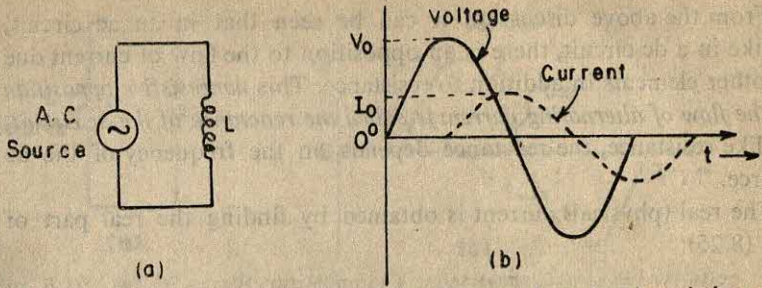


Fig. 8.9 *(a) Ac circuit consisting of an inductor L, and (b) variation of voltage V and current I with time t—notice that the current lags behind the voltage by a phase angle of π/2*

current will also change with time. Let I be the current at a certain instant of time and let dI/dt be the rate of change of current.

Now we know that due to time-varying current, a back (induced) voltage is set up across the inductor.

$$V_{\text{ind}} = -L\frac{dI}{dt}$$

To maintain the flow of current the impressed voltage V must just overcome the back voltage V_{ind}, i.e.

$$V = +L\frac{dI}{dt}$$

Using the complex voltage representation $V = V_0 \exp(i\omega t)$ we have

$$V_0 \exp(i\omega t) = L\frac{dI}{dt}$$

or

$$\frac{dI}{dt} = \frac{V_0}{L}\exp(i\omega t)$$

Integrating, we have

$$I = \frac{V_0}{i\omega L}\exp(i\omega t) = \frac{V}{i\omega L} \tag{8.24}$$

or

$$\frac{V}{I} = i\omega L \tag{8.25}$$

If we compare Eq. (8.25) with Ohm's law [Eq. (8.23)] we find that the quantity $i\omega L$ has the same units as resistance R and may be thought of as the effective resistance offered by the inductor to the flow of current. This quantity is called the *inductive reactance* (or the reactance of the inductor) and is denoted by the symbol

$$Z_L = i\omega L \tag{8.26}$$

From the above discussion it can be seen that in an ac circuit, unlike in a dc circuit, there is an opposition to the flow of current due to other elements in addition to resistance. This *nonresistive opposition to the flow of alternating current is called the reactance of the ac circuit.* Unlike resistance, the reactance depends on the frequency of the ac source.

The real (physical) current is obtained by finding the real part of Eq. (8.25):

$$\begin{aligned} R_e(I) &= R_e\left[\frac{V_0}{i\omega L}\exp(i\omega t)\right] \\ &= \frac{V_0}{\omega L}\,\mathrm{Re}\left[\frac{1}{i}(\cos\omega t + i\sin\omega t)\right] \\ &= \frac{V_0}{\omega L}\,\mathrm{Re}\left[-i\cos\omega t + \sin\omega t\right] \\ &= \frac{V_0}{\omega L}\sin\omega t \end{aligned}$$

which may be rewritten as

$$\begin{aligned} I &= \frac{V_0}{\omega L}\cos\left(\omega t - \frac{\pi}{2}\right) \\ &= I_0\cos\left(\omega t - \frac{\pi}{2}\right) \qquad (8.27) \end{aligned}$$

where $I_0 = V_0/\omega L$.

Notice that now the real current I given by Eq. (8.27) and the real voltage $V = V_0 \cos \omega t$ differ in phase by $\pi/2$. In fact, the current lags behind the voltage by a phase angle of $\pi/2$ or 90°. So when the current is zero, the voltage is maximum and vice versa, as shown in Fig. 8.9(b). Hence, *in an ac circuit containing an ideal inductor, the current lags behind the voltage (or the voltage leads the current) by a phase angle of $\pi/2$ or 90°*. This is evident from the imaginary nature of Z_L.

AC Circuit Containing Only a Capacitor

Let us next consider an ac circuit consisting of an ideal capacitance C [Fig. 8.10(a)]. Let q be the charge on the capacitor plates at any instant of time t. The potential difference between the capacitor plates is, by definition, equal to q/C which must obviously be equal to the applied voltage $V = V_0 \exp(i\omega t)$. Thus

$$\frac{q}{C} = V_0 \exp(i\omega t)$$

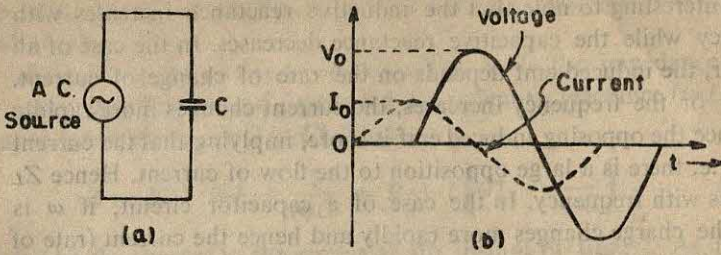


Fig. 8.10 *(a) AC circuit consisting of a capacitance C, and (b) variation of voltage V and current I with time t—notice that the current leads the voltage by a phase angle of π/2*

or
$$q = CV_0 \exp(i\omega t)$$

The current I in the circuit is given by

$$I = \frac{\mathrm{d}q}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}[CV_0 \exp(i\omega t)]$$

$$= CV_0 \frac{\mathrm{d}}{\mathrm{d}t}[\exp(i\omega t)]$$

or
$$I = i\omega CV_0 \exp(i\omega t) = i\omega CV \tag{8.28}$$

or
$$\frac{V}{I} = \frac{1}{i\omega C} \tag{8.29}$$

Comparing Eq. (8.29) with Eq. (8.23) we find that a capacitor in an AC circuit has an effective resistance $1/(i\omega C)$. This is called *capacitative reactance* and is denoted by Z_C. Thus

$$Z_C = \frac{1}{i\omega C} \tag{8.30}$$

The imaginary nature of Z_C once again implies that the current and voltage are not in phase. From Eq. (8.28) the real current is given by

$$\begin{aligned}\mathrm{Re}(I) &= i\omega CV_0 \,\mathrm{Re}\,[\exp(i\omega t)]\\ &= i\omega CV_0 \,\mathrm{Re}\,(\cos\omega t + i\sin\omega t)\\ &= \omega CV_0 \,\mathrm{Re}\,(i\cos\omega t - \sin\omega t)\\ &= -\,\omega CV_0 \sin\omega t\end{aligned}$$

or
$$I = \omega CV_0 \cos\left(\omega t + \frac{\pi}{2}\right)$$

$$= I_0 \cos\left(\omega t + \frac{\pi}{2}\right) \tag{8.31}$$

Notice that the current *leads* the voltage by a phase angle of $\pi/2$ or 90° [Fig. 8.10(b)].

It is interesting to note that the inductive reactance increases with frequency while the capacitive reactance decreases. In the case of an inductor, the induced emf depends on the rate of change of current. So as ω or the frequency increases, the current changes more rapidly and hence the opposing induced emf is more, implying that the current is less, i.e. there is a large opposition to the flow of current. Hence Z_L increases with frequency. In the case of a capacitor circuit, if ω is large, the charge changes more rapidly and hence the current (rate of flow of charge) is large, implying that there is less opposition to the flow of current. Hence, Z_C decreases with frequency.

The LCR Circuit

Finally we consider a very useful ac circuit shown is Fig. 8.11, consisting of all the three elements, namely a resistor R, an inductor L and a capacitor C. The series combination of L, C and R is connected across an ac source.

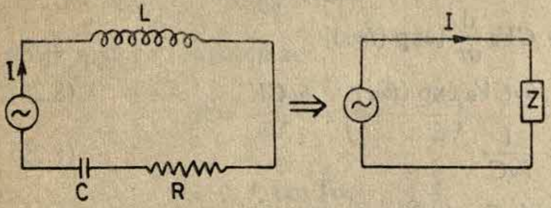


Fig. 8.11 *A series LCR circuit; the series combination of L, C and R is equivalent to total impedance Z*

We have seen that inductors and capacitors in ac circuits are similar to resistors except that their effective resistances are frequency-dependent and that they introduce a phase difference of $\pi/2$ between the voltage and current. The advantage of using complex exponential functions to describe the voltage, current and the effect of inductors and capacitors (i.e. their reactance) is that both the magnitude as well as the phase informations are correctly contained in the functions. To obtain the actual (real) voltage or current, all that we have to do is extract the real part from the complex voltage or current. Furthermore, using the complex variables to describe R, L and C, we can apply to any ac circuit all the methods used earlier for resistors in dc circuits. For example, the rules of series and parallel combinations of resistors and Kirchhoff's laws can also be used. The effective resistance of an AC circuit containing any two or all three elements R, C and L is called *impedance* which is a complex quantity; its real part gives the resistance and the imaginary part is the reactance, as we see in the following.

Referring to the ac circuit shown in Fig. 8.11, the impedance Z is a series combination of resistance R, inductive reactance $(i\omega L)$ and capacitative reactance $(1/i\omega C)$. Thus,

$$Z = R + i\omega L + \frac{1}{i\omega C} = R + i\left(\omega L - \frac{1}{\omega C}\right) \tag{8.32}$$

Let I be the current in the circuit at any instant of time t. As in the case of resistors in series, the same current I flows through the circuit elements L, C and R in series. The current I is given by

$$I = \frac{V}{Z} = \frac{V_0 \exp(i\omega t)}{R + i\left(\omega L - \frac{1}{\omega C}\right)} \tag{8.23}$$

Setting $\left(\omega L - \frac{1}{\omega C}\right) = S$, the physical current (i.e. the real part of I) is

$$\begin{aligned}
\mathrm{Re}(I) &= \mathrm{Re}\left[\frac{V_0 \exp(i\omega t)}{R + iS}\right] \\
&= \mathrm{Re}\left[\frac{V_0(\cos\omega t + i\sin\omega t)\times(R - iS)}{(R + iS)(R - iS)}\right] \\
&= \mathrm{Re}\left[\frac{V_0(R\cos\omega t + S\sin\omega t) + iV_0(R\sin\omega t - S\cos\omega t)}{(R^2 + S^2)}\right] \\
&= \frac{V_0}{(R^2 + S^2)}(R\cos\omega t + S\sin\omega t)
\end{aligned} \tag{8.33}$$

To obtain the phase of Re(I) in relation to that of $V = V_0 \cos\omega t$, we must recast $(R\cos\omega t + S\sin\omega t)$ in Eq. (8.33) as a single cosine function. This is done as follows.

Let $$R = T\cos\delta$$

and $$S = T\sin\delta$$

which give $$T^2 = R^2 + S^2 \text{ or } T = (R^2 + S^2)^{1/2} \tag{8.34}$$

and $$\tan\delta = \frac{S}{R} \tag{8.35}$$

In terms T and δ, we have

$$\begin{aligned}
R\cos\omega t + S\sin\omega t &= T\cos\delta\cos\omega t + T\sin\delta\sin\omega t \\
&= T\cos(\omega t - \delta) \\
&= (R^2 + S^2)^{1/2}\cos(\omega t - \delta)
\end{aligned}$$

where T is given by Eq. (8.34) and δ by Eq. (8.35). Thus Eq. (8.33) may be reacast as

$$\mathrm{Re}(I) = \frac{V_0}{(R^2 + S^2)^{1/2}} \cdot \cos(\omega t - \delta)$$

or

$$\mathrm{Re}(I) = \frac{V_0}{\left[R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2\right]^{1/2}} \cdot \cos(\omega t - \delta)$$

$$= I_0 \cos(\omega t - \delta) \tag{8.36}$$

where

$$I_0 = \frac{V_0}{\left[R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2\right]^{1/2}} = \frac{V_0}{Z} \tag{8.37}$$

and from Eq. (8.35), the phase angle δ is given by

$$\tan \delta = \frac{S}{R} = \frac{\left(\omega L - \frac{1}{\omega C}\right)}{R} \tag{8.38}$$

where δ is the phase difference between the current and voltage.

Variation of I_0 *with* ω It follows from Eq. (8.37) that the amplitude of current I_0, for given L, C and R, varies with angular frequency and that I_0 goes to zero both for $\omega \to 0$ and $\omega \to \infty$. The maximum value of I_0 is reached when the denominator in Eq. (8.37) is minimum, i.e. when

$$\omega L - \frac{1}{\omega C} = 0$$

or

$$\omega = \frac{1}{\sqrt{LC}} \tag{8.39}$$

Then

$$(I_0)_{\max} = \frac{V_0}{R}$$

Figure 8.12 shows the variation of I_0 with ω as ω is increased from low values to high values. Notice that at a particular characteristic frequency $\omega_r = 1/\sqrt{LC}$, I_0 rises to a maximum, depending on R.

The angular frequency ω_r at which I_0 shows a peak is called the "natural" or "resonant" frequency of the ac circuit. Thus we have an electrical analogue of the phenomenon of resonance we have studied earlier in mechanics.

Variation of δ *with* ω It follows from Eq. (8.38) that the phase of the current with respect to that of the voltage, for given L, C and R, depends on the angular frequency ω. In terms of the resonant frequency

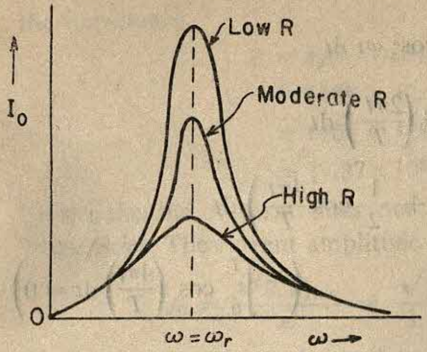


Fig. 8.12 *Variation of the current amplitude with frequency in an LCR circuit*

$\omega_r = 1/\sqrt{LC}$, Eq. (8.38) becomes

$$\tan \delta = \frac{L}{\omega R}(\omega^2 - \omega_r^2)$$

If $\omega < \omega_r$, $\tan \delta$ is negative which means δ is negative implying that the current leads the voltage. If $\omega > \omega_r$, δ is positive indicating that the current lags behind the voltage. But if $\omega = \omega_r$, $\delta = 0$ which means that the current and voltage are in phase if the frequency of the ac source equals the resonant frequency of the ac circuit. The reason why the current amplitude becomes maximum at $\omega = \omega_r$ is that the power delivered to the circuit by the source is maximum at $\omega = \omega_r$ as a result of current and voltage being in the same phase, as we will see in Sec. 8.5.

EXAMPLE 8.5 A coil of inductance 0.50 H and resistance 100 Ω is connected to a 240 V, 50 Hz ac supply.

(a) What is the maximum current in the coil?
(b) What is the time lag between the voltage maximum and current maximum?

[Note: Voltage specification of an ac supply refers to its rms value.]

Solution: It follows from Eqs (8.36) and (8.38) that for an *LR* circuit driven by an ac voltage $V = V_0 \cos \omega t$

$$I = \frac{V_0}{(R^2 + \omega^2 L^2)^{1/2}} \cos(\omega t - \delta) = I_0 \cos(\omega t - \delta)$$

and $$\tan \delta = \omega L/R$$

where V_0 is the rms value of the voltage. The rms value of the alternating voltage is given by ($T = 2\pi/\omega$)

$$(V_{\text{rms}})^2 = \frac{1}{T}\int_0^T V^2\,dt$$

$$= \frac{1}{T}\int_0^T V_0^2 \cos^2 \omega t \, dt$$

$$= \frac{V_0^2}{T}\int_0^T \cos^2\left(\frac{2\pi t}{T}\right) dt$$

$$= \frac{V_0^2}{T}\int_0^T \left(\frac{1}{2} + \frac{1}{2}\cos\frac{4\pi t}{T}\right) dt$$

$$= \frac{V_0^2}{2} \qquad \left(\because \int_0^T \cos\left(\frac{4\pi t}{T}\right) dt = 0\right)$$

$$\therefore \qquad V_{\text{rms}} = \frac{V_0}{\sqrt{2}}$$

or $\qquad V_0 = \sqrt{2}\, V_{\text{rms}}$

(a) It is given that $V_{\text{rms}} = 240$ V, $L = 0.50$ H and $R = 100\ \Omega$. Now,

$$I_0 = \frac{V_0}{(R^2 + \omega^2 L^2)^{1/2}} = \frac{\sqrt{2}\, V_{\text{rms}}}{(R^2 + \omega^2 L^2)^{1/2}}$$

Substituting the numerical values and solving, we get

$$I_0 = 1.82 \text{ A}$$

(b) Since $V = V_0 \cos \omega t$, V is maximum at $t = 0$. Since $I = I_0 \cos(\omega t - \delta)$, I will be maximum when $\omega t - \delta = 0$ or when $t = \delta/\omega$ so that time lag between the voltage maximum and the current maximum is $\Delta t = \delta/\omega$. Now δ is given by

$$\tan \delta = \frac{\omega L}{R} = \frac{2\pi \times 50 \times 0.5}{100} = 1.57$$

which gives $\qquad \delta \simeq 57.5° = 57.5 \times \frac{\pi}{180}$ rad

$$\therefore \qquad \text{Time lag } \Delta t = \frac{\delta}{\omega} = \frac{57.5 \times \pi}{180} \times \frac{1}{(2\pi \times 50)} = 3.2 \times 10^{-3} \text{ s}$$

Example 8.6 Obtain the answers to (a) and (b) in Example 8.5 above if the circuit is connected to a high frequency supply (240 V, 10 kHz). Hence explain the statement that at very high frequency, an inductor in a circuit nearly amounts to an open circuit. How does an inductor behave in a dc circuit after the steady state?

Solution: (a) For the high frequency,

$$\omega = 2\pi \times 10 \text{ kHz} = 2\pi \times 10^4 \text{ rad s}^{-1},$$

the impedance

$$\begin{aligned} Z &= (R^2 + \omega^2 L^2)^{1/2} \\ &= [10^4 + (2\pi \times 10^4)^2 \times 0.5]^{1/2} \\ &= (10^4 + 9.87 \times 10^8)^{1/2} \\ &\simeq (9.87 \times 10^8)^{1/2} \simeq 3.14 \times 10^4\ \Omega \end{aligned}$$

Notice that the R term does not contribute to impedance at high frequencies. The current amplitude is

$$I_0 = \frac{\sqrt{2}\, V_{\mathrm{rms}}}{Z} = \frac{\sqrt{2} \times 240}{3.14 \times 10^4} \simeq 1.1 \times 10^{-2}\ \mathrm{A}$$

Since the current is much smaller now than that for the low frequency case (Example 8.5), a pure inductor at high frequencies behaves as though it were on a nearly open circuit. In a dc circuit when the voltage and current have attained steady values, $\omega = 0$. So the inductor acts like a pure conductor.

(b) At high frequencies, $\omega L > R$, the tangent of phase angle δ becomes very large, i.e. $\tan \delta \gg 1$ which means that $\delta \simeq \pi/2$ rad or 90°. In fact, in this case,

$$\tan \delta = \frac{\omega L}{R} = \frac{2\pi \times 10^4 \times 0.5}{100} = 100\pi$$

which implies that $\delta \simeq \pi/2$ radian.

EXAMPLE 8.7 A 100 μF capacitor in series with a 40 Ω resistance is connected to a 110 V, 60 Hz supply.

(a) What is the maximum current in the circuit?

(b) What is the time lag between the current maximum and voltage maximum?

(The voltage of the ac supply refers to its rms value)

Solution: It follows from Eqs (8.36) and (8.38) that for a CR circuit driven by an ac voltage $V = V_0 \cos \omega t$

$$I = \frac{V_0}{\left(R^2 + \dfrac{1}{\omega^2 C^2}\right)^{1/2}} \cos(\omega t - \delta)$$

and

$$\tan \delta = -\frac{1}{\omega CR}$$

(a) It is given that $V_{\mathrm{rms}} = 110$ V, $\omega = 2\pi \times 60$ rads^{-1}, $R = 40\ \Omega$ and $C = 100\ \mu\mathrm{F} = 10^{-4}$ F. Now

$$I_0 = \frac{\sqrt{2}\, V_{\mathrm{rms}}}{\left(R^2 + \dfrac{1}{\omega^2 C^2}\right)^{1/2}}$$

Substituting the values and solving, we get

$$I_0 = 3.24 \text{ A}$$

(b) The phase angle δ is given by

$$\tan\delta = -\frac{1}{\omega CR} = -\frac{1}{(2\pi \times 60) \times 10^{-4} \times 40}$$
$$= -0.6631$$

giving $$\delta = -33.5° = -33.5 \times \frac{\pi}{180} \text{ rad}$$

Time lag $$\Delta t = \frac{\delta}{\omega} = -33.5 \times \frac{\pi}{180} \times \frac{1}{2\pi \times 60} = -1.55 \times 10^{-3} \text{ s}$$

The minus sign indicates that the current leads the voltage.

Example 8.8 Obtain the answers to (a) and (b) in Example 8.7 above if the circuit is connected to a 110 V, 12 kHz supply? Hence explain the statement that a capacitor is a conductor at very high frequencies. Compare this behaviour with that of a capacitor in a dc circuit after the steady state?

Solution: (a) For the high frequency, $\omega = 2\pi \times 12 \times 10^3$ rad s^{-1}, the impedance is

$$Z = \left(R^2 + \frac{1}{\omega^2 C^2}\right)^{1/2} \simeq R,$$

since for $C = 10^{-4}$ F and $R = 40\ \Omega$, $1/\omega C \ll R$, so that the current amplitude is

$$I_0 = \frac{\sqrt{2} \times 110}{40} = 3.9 \text{ A}$$

Notice that the current I_0 is now greater than that in the low frequency case (Example 8.7). This means that, at high frequencies, a capacitor conducts more charge than at low frequencies, i.e. it behaves as a conductor. This is surprising. How can a capacitor conduct current even when there is nothing (air) between its plates? The answer to this question will be given in the next chapter when we will learn about a new kind of current called 'displacement current' which will be shown to exist even in a vacuum.

(b) $$\tan\delta = -\frac{1}{\omega CR} = \frac{1}{(2\pi \times 12 \times 10^3) \times (10^{-4}) \times (40)} = -\frac{1}{96\pi}$$

which has a very small negative value. Thus tan δ → minus zero, which means that, at high frequencies, the phase difference between current and voltage is nearly π rad or 180°. The negative sign shows that the current leads voltage at high frequencies.

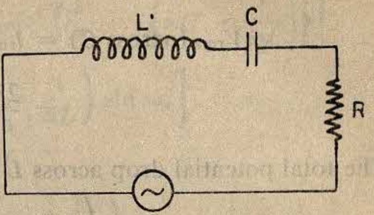


Fig. 8.13

EXAMPLE 8.9 Figure 8.13 shows a series LCR circuit connected to a variable frequency 230 V source. $L = 5.0$ H, $C = 80\ \mu$F, and $R = 40\ \Omega$.

(a) Determine the source frequency which drives the circuit in resonance.

(b) Obtain the impedance of the circuit and the amplitude of current at the resonating frequency.

(c) Determine the rms potential drops across the three lements of the circuits. Show that the potential drops across the LC combination is zero at the resonating frequency.

Solution: (a) The resonant angular frequency is

$$\omega_r = \frac{1}{\sqrt{LC}} = \frac{1}{(5.0\times 80\times 10^{-6})^{1/2}} = 50 \text{ rad s}^{-1}$$

Therefore, the resonant frequency is

$$\nu_r = \frac{\omega_r}{2\pi} = \frac{50}{2\pi} = \frac{25}{\pi}\text{ Hz}$$

(b) The impedance is given by

$$Z = \left[R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2\right]^{1/2}$$

When $\omega = \omega_r = \frac{1}{\sqrt{LC}}$ (i.e. at resonance), $\omega L = \frac{1}{\omega C}$, and therefore

$$Z = R = 40\ \Omega$$

Current amplitude at resonance is

$$I_0 = \frac{V_0}{Z} = \frac{V_0}{R} = \frac{\sqrt{2}\,V_{\text{rms}}}{R} = \frac{\sqrt{2}\times 230}{40} = 8.1 \text{ A}$$

Also $\quad I = \frac{V_{\text{rms}}}{R} = \frac{230}{40} = \frac{23}{4}$ A

(c) The rms potential drop across L at resonance is

$$V_{\text{rms}}\text{ (across } L) = I_{\text{rms}}\times\omega_r L$$

$$= \frac{23}{4}\times 50\times 5 = 1437.5 \text{ V}$$

The rms potential drop across C at resonance is

$$V_{\mathrm{rms}}\ (\text{across } C) = I_{\mathrm{rms}} \times \frac{1}{\omega_r C}$$

$$= \frac{23}{4} \times \frac{1}{50 \times 80 \times 10^{-6}} = 1437.5 \text{ V}$$

The total potential drop across L and C is (at resonance)

$$I_{\mathrm{rms}} \times \left(\omega_r L - \frac{1}{\omega_r C}\right) = 0$$

The rms potential drop across R at resonance is

$$V_{\mathrm{rms}}\ (\text{across } R) = I_{\mathrm{rms}} \times R$$

$$= \frac{23}{4} \times 40 = 230 \text{ V}$$

$$= \text{applied rms voltage}$$

This is expected since the total potential drop across L and C is zero.

EXAMPLE 8.10 Keeping the source frequency equal to the resonating. frequency of the series LCR circuit, if the three elements L, C and R are arranged in parallel, show that the total current in the parallel LCR circuit is a minimum at this frequency. Obtain the current rms value in each branch of the circuit for the elements and source specified in Example 8.9 for this frequency.

Solution: Figure 8.14 shows the parallel LCR circuit.

The impedance of the circuit with L, C and R connected in parallel is obtained by using the rule of a parallel combination of resistors. Since reactances are $i\omega L$ (inductive) and $1/i\omega C$ (capacitative), the impedance Z of the parallel combination is given by

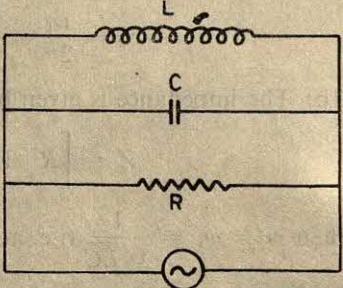


Fig. 8.14 *A parallel LCR circuit*

$$\frac{1}{Z} = \frac{1}{R} + \frac{1}{i\omega L} + i\omega C$$

$$= \frac{1}{R} + i\left(\omega C - \frac{1}{\omega L}\right)$$

The current is

$$I = \frac{V}{Z} = V_0 \exp(i\omega t) \times \left[\frac{1}{R} + i\left(\omega C - \frac{1}{\omega L}\right)\right]$$

Now

$$\begin{aligned}\text{Re}(I) &= V_0 \,\text{Re}\left[(\cos \omega t + i \sin \omega t)\left\{\frac{1}{R} + i\left(\omega C - \frac{1}{\omega L}\right)\right\}\right]\\ &= V_0\left[\frac{1}{R}\cos \omega t - \left(\omega C - \frac{1}{\omega L}\right)\sin \omega t\right]\\ &= V_0(a \cos \omega t - b \sin \omega t) \qquad \text{(i)}\end{aligned}$$

where $a = 1/R$ and $b = \left(\omega C - \frac{1}{\omega L}\right)$

Equation (i) can be recast in the form of a single cosine function if we set

$$a = r \cos \delta$$

and

$$b = r \sin \delta$$

where $r = (a^2 + b^2)^{1/2}$ and $\tan \delta = b/a$. In terms of r and δ, Eq. (i) reads

$$\begin{aligned}\text{Re}(I) &= V_0 r \cos(\omega t + \delta)\\ &= \frac{V_0}{Z}\cos(\omega t + \delta) \qquad \text{(ii)}\end{aligned}$$

where impedance Z is given by

$$Z = \frac{1}{r} = \frac{1}{(a^2 + b^2)^{1/2}} = \frac{1}{\left[\frac{1}{R^2} + \left(\omega C - \frac{1}{\omega L}\right)^2\right]^{1/2}} \qquad \text{(iii)}$$

At resonant frequency $\omega_r = \frac{1}{\sqrt{LC}}$, $\omega C = \frac{1}{\omega L}$ and the denominator in Eq. (iii) is the minimum so that impedance Z is the maximum. Hence, at resonance, the current in the circuit is the minimum. Also at resonance, $Z = R$.

$$I_{\text{rms}}\text{ (in circuit)} = \frac{V_{\text{rms}}}{R} = \frac{230}{40} = 5.75 \text{ A}$$

The rms current in the R branch is

$$I_{\text{rms}}\text{ (in } R) = \frac{V_{\text{rms}}}{R} = 5.75 \text{ A}$$

The rms current in L branch is

$$I_{\text{rms}}\text{ (in } L) = \frac{V_{\text{rms}}}{\omega_r L} = \frac{230}{50 \times 5.0} = 0.92 \text{ A}$$

The rms current in C branch is

$$\begin{aligned}I_{\text{rms}}\text{ (in } C) &= V_{\text{rms}} \times \omega_r C = 230 \times 50 \times 80 \times 10^{-6}\\ &= 0.92 \text{ A}\end{aligned}$$

Note that, at resonance, the total current in the circuit = current through R only. The reason is that, at resonance, the currents in L and C branches are 180° out of phase (i.e. they are equal in magnitude but opposite in sign) and add up to zero at every instant of time during the cycle.

8.5 ENERGY AND POWER ASSOCIATED WITH L, C AND R

In Sec. 8.2 we have seen that in an inductor it takes time for current to grow from zero to a steady value I_0. Let I be the current at any instant of time t during the growth of the current. The inductor produced a back emf $\mathcal{E} = -L\,\mathrm{d}I/\mathrm{d}t$ which tends to prevent this growth of current. To maintain a growth of current, power equal to $\mathcal{E}I$ has to be supplied from outside (external source). Hence the total amount of work done by the external source in building up the current from an initial value zero to a final value I_0 is

$$W = \int_0^{I_0} L\frac{\mathrm{d}I}{\mathrm{d}t}\,\mathrm{d}t = L\int_0^{I_0} I\,\mathrm{d}I = \frac{1}{2}LI_0^2 \tag{8.40}$$

If the external source happens to be an ac source supplying an ac voltage $V = V_0 \cos \omega t$, then the current through L builds up from zero to a maximum value $I_0(= V_0/\omega L)$ and then falls back to zero [see Eq. (8.27)]. The source supplies energy equal to $\frac{1}{2}LI_0^2$ during the build-up of the current from 0 to I_0 and this energy is returned back to the source during the reverse process (i.e. during the fall of current from I_0 to zero) so that no net power is supplied by the source in a complete cycle. This can also be shown as follows ($T = 2\pi/\omega$)

$$W = \int_0^T VI\,\mathrm{d}t$$

Now $V = V_0 \cos \omega t$ and $I = I_0 \cos\left(\omega t - \frac{\pi}{2}\right) = I_0 \sin \omega t$ [see Eq. (8.27)]. Therefore,

$$W = V_0I_0 \int_0^T \cos \omega t \cdot \sin \omega t\,\mathrm{d}t$$

$$= \frac{V_0I_0}{2}\int_0^T \sin(2\omega t)\,\mathrm{d}t$$

$$= \frac{V_0I_0}{2}\int_0^T \sin\left(\frac{4\pi t}{T}\right)\mathrm{d}t$$

or

$$W = 0 \tag{8.41}$$

since the integrand has the same value (= zero) at the two limits.

Exactly the same thing happens when a capacitor is charged and discharged. We know that the electrostatic energy associated with a capacitor of capacitance C with a potential difference V across its plates is $\frac{1}{2}CV^2$. The external source has to supply this energy to charge it, but the energy is returned to the source when the capacitor is discharged. If the external source is an ac source connected across the capacitor, it is charged during half the cycle and discharged during the next half so that the net energy during a complete cycle is zero. This can be seen as follows:

The total energy supplied during a time period T is mathematically given by [refer to Eq. (8.31)]

$$\begin{aligned} W &= \int_0^T VI\,\mathrm{d}t = \int_0^T V_0 \cos \omega t \cdot I_0 \cos\left(\omega t + \frac{\pi}{2}\right) \cdot \mathrm{d}t \\ &= -V_0 I_0 \int_0^T \cos \omega t \cdot \sin \omega t \cdot \mathrm{d}t \\ &= 0 \qquad \text{as before} \end{aligned} \tag{8. 42}$$

The reason for the net energy in both the cases studied above being zero is that the current and voltage have a phase difference of $\pm\pi/2$ [see Eqs (8.27) and (8.31)].

The situation in the case of a resistor is very different. For a resistor, the current and voltage are always in phase [see Eq. (8.22)]. Hence, although the current in a resistor (connected across an ac source) oscillates just as it does in an inductor or a capacitor, the power (i.e. product of voltage and current) is always *positive*. The average power consumed in a full cycle of period T is [see Eq. (8.22)]

$$\begin{aligned} P = \frac{W}{T} &= \frac{1}{T}\int_0^T VI\,\mathrm{d}t = \frac{V_0 I_0}{T}\int_0^T \cos^2 \omega t\,\mathrm{d}t \\ &= \frac{V_0 I_0}{T}\int_0^T \left(\frac{1}{2} + \frac{1}{2}\cos 2\omega t\right)\mathrm{d}t \\ &= \frac{V_0 I_0}{T}\left(\frac{T}{2} + 0\right) \\ &= \frac{V_0 I_0}{2} \\ &= \frac{V_0^2}{2R} \qquad (\because \quad V_0 = RI_0) \end{aligned} \tag{8.43}$$

Equation (8.43) is often expressed in terms of rms values as

$$V_{\text{rms}}^2 = \frac{1}{T}\int_0^T V_0^2 \cos^2 \omega t \cdot \mathrm{d}t = \frac{V_0^2}{2}$$

$$I_{\text{rms}}^2 = \frac{1}{T}\int_0^T I_0^2 \cos^2 \omega t\cdot \mathrm{d}t = \frac{I_0^2}{2}$$

Therefore $$P = \frac{1}{2}V_0 I_0 = V_{\text{rms}}\cdot I_{\text{rms}} \tag{8.44}$$

The Choke Coil—A Simple Useful Device

A choke is simply an inductor (coil) having a large self-inductance. The usefulness of such a simple device stems from the fact that, if fed by an ac source, it does not consume any power [as we have seen in Eq. (8.41)]. It is connected in series with an electrical appliance of resistance, say, R (such as a fluorescent tube) and the combination is connected across an ac source as shown in Fig. 8.15. We have already studied a more complicated series circuit in which in addition to L and R, we had a capacitor C in series as well (Fig. 8.11).

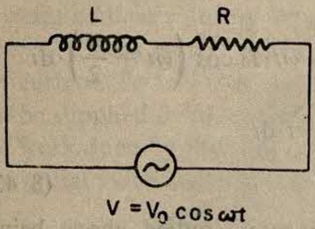


Fig. 8.15

We can, therefore, use the results (8.32) and (8.36) obtained earlier with the quantity C absent from them. The expressions now read:

$$Z = R + i\omega L$$

with $$\text{Re}\,(I) = I_0 \cos(\omega t - \delta)$$

$$I_0 = \frac{V_0}{(R^2 + \omega^2 L^2)^{1/2}}$$

and

$$\tan\delta = \frac{\omega L}{R}$$

The voltage across R is

$$\begin{aligned} V_R &= \text{Re}\,(I)R \\ &= \frac{V_0 R}{(R^2+\omega^2L^2)^{1/2}}\cdot\cos(\omega t - \delta) \\ &= V_{\text{m}}\cos(\omega t - \delta) \end{aligned}$$

where V_{m} is the amplitude of V_R (i.e. maximum value of V_R)

$$V_{\text{m}} = \frac{V_0 R}{(R^2+\omega^2L^2)^{1/2}}$$

(The student is advised to obtain these results from first principles.)

Notice that $V_m < V_0$, the amplitude of the voltage across R in the absence of L. Thus, by choosing an appropriate value of L, the current through R and voltage across R can be made smaller than the values obtained with $L = 0$, to any desired value. We have already seen that no power is lost in the inductor and in reducing the voltage across R, we do not waste any power. We could, of course, reduce voltage across R by connecting an additional resistor in series with it, but this would involve wastage of power.

The choke is used in series with a fluorescent tube working on ac mains. If the tube is connected directly across 220 V, a large current would flow through it and the tube may get damaged. The use of a choke in series with the tube decreases the voltage across the tube to a safe low value without any waste of power.

Electrical Resonance

In Sec. 8.4 we learnt that a series LCR circuit driven by an ac voltage $V = V_0 \cos \omega t$ shows resonance when $\omega = \omega_r = 1/\sqrt{LC}$. We have seen that if the frequency ω of the applied voltage is equal to the natural frequency ω_r, the impedance of the circuit becomes minimum [see Eq. (8.37)] and consequently the current amplitude becomes maximum [see Eq. (8.37) and Fig. 8.12)]. The reason is that, at resonance, i.e. when $\omega = \omega_r = 1/\sqrt{LC}$, the circuit extracts maximum power from the source. This can be proved as follows.

The average power transferred to the circuit by the ac source in one full cycle, i.e. during time $T = 2\pi/\omega$, is

$$P = \frac{1}{T}\int_0^T \mathrm{Re}\,(V)\,\mathrm{Re}\,(I)\,\mathrm{d}t$$

$\mathrm{Re}\,(V) = V_0 \cos \omega t$ and $\mathrm{Re}\,(I)$ is given by Eq. (8.36). Hence

$$\begin{aligned} P &= \frac{1}{T}\int_0^T (V_0 \cos \omega t)[I_0 \cos(\omega t - \delta)]\,\mathrm{d}t \\ &= \frac{V_0 I_0}{T}\int_0^T \cos \omega t\,(\cos \omega t \cos \delta + \sin \omega t \sin \delta)\,\mathrm{d}t \\ &= \frac{V_0 I_0}{T}\left[\cos \delta \int_0^T \cos^2 \omega t \cdot \mathrm{d}t + \sin \delta \int_0^T \sin \omega t \cos \omega t\,\mathrm{d}t\right] \\ &= \frac{V_0 I_0}{T}\left(\cos \delta \cdot \frac{T}{2} + 0\right) \end{aligned}$$

or $$P = \frac{1}{2} V_0 I_0 \cos \delta \tag{8.45}$$

where I_0 and δ are given by Eqs. (8.37) and (8.38):

$$I_0 = \frac{V_0}{\left[R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2\right]^{1/2}} \tag{8.46}$$

and

$$\tan \delta = \frac{\omega L - \frac{1}{\omega C}}{R} \tag{8.47}$$

Now $\cos \delta = 1/(1 + \tan^2 \delta)^{1/2}$. Substituting for $\tan \delta$ from Eq. (8.47) and simplifying, we get

$$\cos \delta = \frac{R}{\left[R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2\right]^{1/2}} \tag{8.48}$$

Now, using Eqs (8.46) and (8.48) in Eq. (8.45), we have

$$P = \frac{1}{2} \frac{V_0^2 R}{\left[R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2\right]} \tag{8.49}$$

At resonance, i.e. when $\omega = 1/\sqrt{LC}$, $\omega L = 1/\omega C$ and the denominator becomes minimum $= R^2$, hence P becomes maximum:

$$P_{\text{max}} = \frac{1}{2} \frac{V_0^2}{R}$$

In terms of P_{max}, Eq. (8.49) reads

$$P = P_{\text{max}} \frac{R^2}{\left[R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2\right]} \tag{8.50}$$

Thus we conclude that when the frequency ω of the applied voltage is equal to the resonant frequency $\omega_r = 1\sqrt{LC}$ of the circuit, the power delivered to the circuit by the source is maximum. Also notice from Eq. (8.49) that $P \to 0$ both when $\omega \to 0$ and $\omega \to \infty$, i.e. at very low and very high frequencies (compared to ω_r) the power input is negligibly small. Figure 8.16 shows the variation of power input P as the frequency of the applied voltage is increased from low to high values. Notice the peak of the power curve at $\omega = \omega_r$.

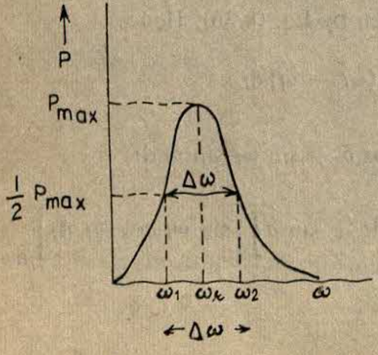


Fig. 8.16 *Variation of P with ω; half power points*

Half Power Points

The values of ω at which the power input (P) is half its maximum value ($P_{\max}$) are called *half power points*. From Eq. (8.50), the half power points are the values of ω satisfying the equation:

$$\frac{1}{2} = \frac{R^2}{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}$$

or $$\omega L - \frac{1}{\omega C} = \pm R$$

or $$\omega^2 - \frac{1}{LC} = \pm \frac{R}{L}\omega$$

or $$\omega^2 - \omega_r^2 = \pm \frac{R}{L}\omega$$

These are two quadratic equations in ω, namely, $\omega^2 + (R/L)\omega - \omega_r^2 = 0$ and $\omega^2 - (R/L)\omega - \omega_r^2 = 0$. Each of these equations has one positive and one negative root. Since angular frequency cannot be negative, we retain only the positive roots which are:

$$\omega_1 = -\frac{R}{2L} + \left(\omega_r^2 + \frac{R^2}{4L^2}\right)^{1/2} \tag{8.51a}$$

and $$\omega_2 = +\frac{R}{2L} + \left(\omega_r^2 + \frac{R^2}{4L^2}\right)^{1/2} \tag{8.51b}$$

These values are indicated in Fig. 8.16. The frequency interval between the two half power points is

$$\Delta\omega = \omega_2 - \omega_1 = \frac{R}{L} \tag{8.52}$$

Thus the value of R/L determines the width of the power curve at half power points (see Fig. 8.16). $\Delta\omega$ is called the full-width at half maximum power or bandwidth.

Sharpness of Resonance

Resonance is said to be sharp if the power-frequency curve of Fig. 8.16 falls steeply around $\omega = \omega_r$. The sharpness of resonance is quantitatively described in terms of what is called the *quality* or *Q factor* defined as:

$$Q = \frac{\text{resonant frequency}}{\text{bandwidth}} = \frac{\omega_r}{\Delta\omega}$$

or

$$Q = \frac{1}{\sqrt{LC}} \cdot \frac{L}{R} = \sqrt{\frac{L}{C}} \cdot \frac{1}{R} \tag{8.53}$$

Q, which is a dimensionless number, is an indicator of the sharpness of the power peak; a higher Q indicating a smaller $\Delta\omega$ and hence a sharper peak.

The most familiar example of resonance in electrical circuits is encountered when we tune our radio receiver set to a particular broadcasting frequency (or station, as we call it). The transmitter at the broadcasting centre sends radiowaves over a band of frequencies. A particular setting of the tuner corresponds to a particular set of values of L and C and hence to a particular frequency of the circuit. When this frequency equals the frequency of the signal from the broadcasting station, resonance takes place and the power absorption becomes maximum and hence we hear only that signal.

EXAMPLE 8.11 A series LCR circuit with $L = 0.12$ H, $C = 480$ nF, $R = 23\ \Omega$ is connected to a 230 V variable frequency supply.

(a) What is the source frequency for which the current amplitude is maximum? Obtain this maximum value.

(b) What is the source frequency for which the average power absorbed by the circuit is maximum? Obtain the value of this maximum power.

(c) For which frequencies of the source is the power transferred to the circuit half the power at resonant frequency? What is the current amplitude at these frequencies?

(d) What is the Q-factor of the given circuit?

Solution: $L = 0.12$ H $\quad C = 480$ nF $= 480 \times 10^{-9}$ F $\qquad R = 23\ \Omega$

$$V_0 = \sqrt{2}\, V_{\text{rms}} = \sqrt{2} \times 230 \text{ V}$$

(a) Current amplitude I_0 is maximum at resonant angular frequency

$$\omega_r = \frac{1}{\sqrt{LC}} = \frac{1}{(0.12 \times 480 \times 10^{-9})^{1/2}} = 4167 \text{ rad s}^{-1}$$

$$\nu_r = \frac{\omega_r}{2\pi} = 663 \text{ Hz}$$

From Eq. (8.46) the maximum value of I_0 is

$$(I_0)_{\max} = \frac{V_0}{R} = \frac{\sqrt{2} \times 230}{23} = 14.1 \text{ A}$$

(b) Average power P given by Eq. (8.49) is maximum at resonant frequency $\nu_r = 663$ Hz.

$$P_{\max} = \frac{1}{2}\frac{V_0^2}{R} = \frac{1}{2}\times\frac{(\sqrt{2}\times 230)^2}{23} = 2300 \text{ W}$$

(c) The two angular frequencies are given by (see Eqs 8.51a and b) ($\because R/2L \leqslant \omega_r$)

$$\omega_1 \simeq \omega_r - \frac{R}{2L}$$

and
$$\omega_2 \simeq \omega_r + \frac{R}{2L}$$

Substituting the numerical values and solving, we get

$$\omega_1 = 4071 \text{ rad s}^{-1} \quad \text{or} \quad \nu_1 = 648 \text{ Hz}$$

$$\omega_2 = 4263 \text{ rad s}^{-1} \quad \text{or} \quad \nu_2 = 678 \text{ Hz}$$

At these frequencies, power absorbed $= \frac{1}{2}P_{\max}$. Since power is proportional to (current amplitude)^2, the current amplitude at these frequencies is $\frac{1}{2}(I_0)_{\max} = 14.1/\sqrt{2} = 10$ A

(d)
$$Q = \frac{\omega_r L}{R} = \frac{4167\times 0.12}{23} = 21.7$$

8.6 OSCILLATIONS IN AN LC CIRCUIT

In class XI you have studied oscillations of mechanical systems. We will now discuss the oscillations of an electrical system consisting of a capacitor C and an inductor L (Fig. 8.17). The equilibrium state is the one in which the capacitor is uncharged and no current is flowing in the circuit. This state is disturbed by pressing the key, thus charging the capacitor. Let q be the charge on the capacitor at any instant of time so that $V = q/C$ is the voltage across the capacitor plates at that instant. When the key is released, the capacitor starts discharging through the inductor, i.e. the charge changes with time and a current $I = dq/dt$ is established in the inductor.

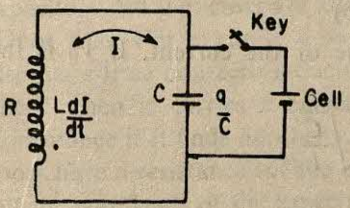


Fig. 8.17 *The oscillatory LC circuit*

In this circuit the restoring force is due to the force of repulsion between electrons. This force tends to distribute electrons equally on the capacitor plates so that there is no net charge. Inductance, on the other hand, tends to oppose this redistribution, i.e. it opposes the increase of current. At any instant of time, the voltage across the inductor is

$$V = -L\frac{dI}{dt} = -L\frac{d^2q}{dt^2}$$

The minus sign indicates that the voltage opposes the increase of current. From Kirchhoff's law, this voltage must equal the voltage q/C across the capacitor plates, giving,

$$-L\frac{d^2q}{dt^2} = \frac{q}{C}$$

or

$$\frac{d^2q}{dt^2} = -\frac{1}{LC}q = -\omega^2 q$$

with

$$\omega = \frac{1}{\sqrt{LC}}$$

Thus, in an electrical circuit consisting of an inductance L and a capacitance C, the charge oscillates harmonically with an angular frequency $\omega = 1/\sqrt{LC}$ and period $T = 2\pi\sqrt{LC}$. At any instant of time, the charge q is given by

$$q = q_0 \sin(\omega t + \delta)$$

where q_0 is the maximum value of the charge and δ is the phase of electron oscillations. The current in the circuit is given by

$$I = \frac{dq}{dt} = \omega q_0 \cos(\omega t + \delta)$$

or

$$I = I_0 \cos(\omega t + \delta)$$

where $I_0 = \omega q_0$ is the maximum value of the current. If V_0 is the applied voltage,

$$I_0 = V_0\sqrt{\frac{C}{L}}$$

since $q_0 = CV_0$ and $\omega = 1/\sqrt{LC}$.

Energy Considerations

- To understand how an LC circuit oscillates, let us assume that initially the capacitor C carries a charge q and the current in inductor L is zero. At this instant, the electrostatic energy stored in the capacitor is

$$E_e = \frac{1}{2}\frac{q^2}{C}$$

and that in the inductance is zero, since $I = 0$ initially. As time passes, the capacitor starts discharging through the inductance and a current $I = \mathrm{d}q/\mathrm{d}t$ is established in the inductor. As q decreases, E_e decreases and I increases, so that the energy now appears around the inductance as the current builds up. When the capacitor is completely discharged, the magnetic energy is

$$E_m = \frac{1}{2}LI^2$$

The total energy of the system is conserved, since the system considered here does not contain any resistive component, so that there is no dissipation of energy. The total energy is

$$\begin{aligned} E &= E_e + E_m \\ &= \frac{1}{2}\frac{q^2}{C} + \frac{1}{2}LI^2 \\ &= \frac{1}{2C}[q_0^2 \sin^2(\omega t + \delta)] + \frac{L}{2}[I_0^2 \cos^2(\omega t + \delta)] \\ &= \frac{1}{2}CV_0^2[\sin^2(\omega t + \delta) + \cos^2(\omega t + \delta)] \\ &\qquad (\because q_0 = CV_0,\ I_0 = \omega q_0,\ \omega^2 = 1/LC) \\ &= \frac{1}{2}CV_0^2 = \text{constant, as expected} \end{aligned}$$

8.7 REAL RESISTORS, CAPACITORS AND INDUCTORS

In our analysis of electrical circuits we have assumed that the resistors, capacitors and inductors are all ideal or pure. In actual practice this is not true. For example, a real metal wire has some resistance but it also has a little non-zero inductance because a magnetic field is produced when it carries a current. It may also have a little non-zero capacitance if it finds another wire near it in a circuit. Similarly capacitors have a resistance (of the dielectric) which is of course very high, but the wire leads of the capacitor have inductance. Also an inductor (which is a wire wound in the form of a coil) will certainly have a certain resistance, and there is some capacitance between individual successive turns of the coil.

SUMMARY

In an ac circuit, the current in a resistor and the voltage across it are always in phase. If the circuit has an inductor, the current in it will lag behind the voltage across it by $\pi/2$. The current in a capacitor leads the voltage by $\pi/2$. Capacitor and inductor offer impedance to the flow of current in an ac circuit.

In an *LCR* circuit, the average power absorption in the inductance and capacitance is zero; it is the pure resistive component which absorbs power. The power absorption is maximum if the frequency of the applied voltage equals the resonant (natural) frequency of the circuit. In an *LC* circuit, charge, current and voltage oscillate harmonically at an angular frequency $\omega = 1/\sqrt{LC}$.

EXERCISES

A. Short-Answer Questions

Pick out "true" and "false" statements from the following.

1. In an inductor, the current rises to a steady value at a constant rate.
2. The time constant of an *RC* circuit is given by R/C.
3. During discharging, the charge on a capacitor decays exponentially with time.
4. In an ac circuit consisting of a pure resistor, the current and voltage are always in phase.
5. In an *LC* circuit, driven by an alternating voltage source, the average power during a cycle delivered by the source is zero.
6. The impedance of an *LC* circuit is independent of the frequency of the applied voltage.
7. The resonant frequency of an *LCR* circuit depends upon the values of *L*, *C* and *R*.
8. A choke is always used in parallel with a fluorescent tube.
9. At resonance, the amplitude of the current rises to a maximum value.
10. A very small value of the quality factor implies a sharp resonance.

Long-Answer Question

1. (a) Describe briefly how the current rises and falls in a dc circuit containing an inductor and resistor. Does the theory given in the text remain valid if the internal resistance of the battery is not negligible?
   (b) Is the direction of current during its growth and decay in the *LR* circuit the same?
   (c) What is time constant? Explain its significance.
   (d) Under what circumstances does an inductor behave as (i) a large resistor, and (ii) a short resistor, i.e. a resistor of zero resistance?
   (e) Explain why there is usually a spark near the switch in a dc circuit containing a large inductor and a resistor when the switch is partially opened. What harm can such a spark cause to the circuit and how can such a spark be prevented?
   (f) Explain why the electrical connections in any circuit have to be non-inductive.
   (g) How are the resistances in a resistance box wound non-inductively.

2. (a) What is a capacitor? Does the capacity of a capacitor depend on the charge on the plate?
   (b) Describe briefly how the charge and current in an *RC* circuit vary when the capacitor is charged and discharged. Does the direction of current in the charging and discharging process remain the same?
   (c) Is the significance of time constant in *LR* and *CR* circuits the same?
   (d) Under what conditions does a capacitor behave as a large resistor?
3. Describe briefly how the charge on the plates of a capacitor oscillates when connected to an ideal inductor.
4. What type of energy is stored in: (a) an inductor, (b) a capacitor, and (c) a secondary cell?
5. A pure resistor is connected across an ac source. (a) Show that the current in the resistor and the voltage across it are always in phase. (b) Deduce the expression for the average power delivered by the source in one complete cycle.
6. A pure inductor is connected across an alternating source. (a) Show that the voltage leads the current by a phase angle of 90°. (b) Find the time-average power delivered by the source.
7. An ideal capacitor is connected across an alternating source. (a) Show that the current leads the voltage by a phase angle of 90°. (b) Show that the time-averaged power delivered by the source is zero.
8. Distinguish clearly between reactance and resistance.
9. A series combination of *L*, *C* and *R* is connected across an ac voltage $V = V_0 \sin \omega t$. Find the impedance of the circuit. Obtain an expression for the current. Discuss the variation of the amplitude and phase of the current with the frequency of the applied voltage.
10. Answer Q.9 if *L*, *C* and *R* are connected in parallel.
11. What is a choke? How is it useful?
12. What is electrical resonance of an *LCR* circuit. Show that the time-averaged power transferred by the source to the circuit is maximum at resonance. What is the maximum power?
13. Define half-power points and quality factor. Discuss the sharpness of resonance in relation to an *LCR* circuit driven by an alternating voltage.
14. (a) Define the rms value of current for an alternating current.
    (b) Why is there a need to define rms voltage and current?
    (c) What is the relationship between maximum current and rms current? Does this relation hold for any type of variation of current with time?
    (d) Can such currents be measured with an ordinary galvanometer?
15. Describe briefly how alternating currents are measured.
16. (a) The source of alternating potential has some impedance. When such a source is connected to *L*, *C* and *R*, is there any restriction on the value of impedance of source as compared to impedance of the circuit?
    (b) Is the instantaneous value of applied voltage equal to the sum of the instantaneous voltage drops across each element of the circuit?
17. (a) An air core coil and a bulb are connected to an ac mains supply (220 V, 50 Hz). If an iron rod is inserted in the coil, does the bulb become more and more dim as more and more of the iron rod is inserted? Explain. Mention one place where such a device finds practical application. Can the same behaviour be observed when dc is used instead of ac?

(b) If the inductor is replaced by a capacitor, will the bulb still become dim as the value of the capacitor is increased?

(c) An ac source is connected to a bulb, an inductor and a capacitor in series; explain how such an arrangement can be used so that there is a normal brilliance in the bulb without the inductor and the capacitor. Both inductive and capacitive reactances represent opposition to the flow of charge; one decreases with frequency while the other increases with it. Explain physically how it happens.

18. In the case of a dc circuit, the power loss across a resistor is equal to the product of current and voltage. In an ac circuit also the power loss across a resistor is the product of the effective current and effective voltage. Can this definition be extended to an ideal inductor or capacitor? If yes, what is the power spent in an inductor?

C. Multiple-Choice Questions

Choose the correct answer from the given aternatives:

1. In the circuit shown in Fig. 8.18, the key K is plugged at time $t = 0$. Which one of the graphs shown in Fig. 8.19 represents the variation of current I in the circuit with time t.

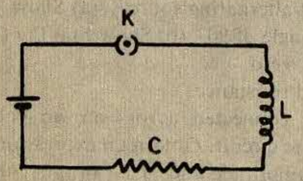
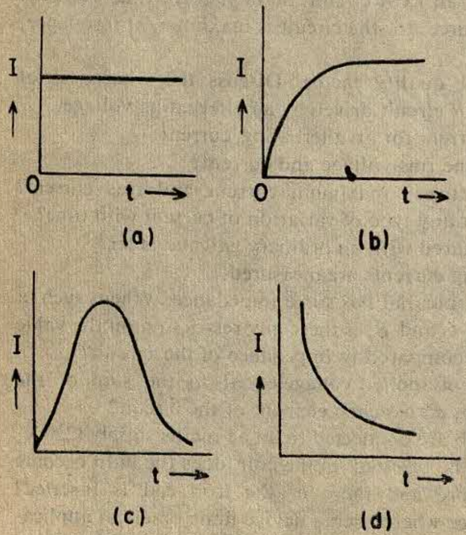


Fig. 8.18

2. Which one of the graphs shown in Fig. 8.20 represents the variation of voltage V across inductor L with time t in Fig. 8.18?

3. Two resistors of 10 Ω and 20 Ω and an ideal inductor of inductance 5 H are connected to a 2 V battery



(a) (b) (c) (d)

Fig. 8.19

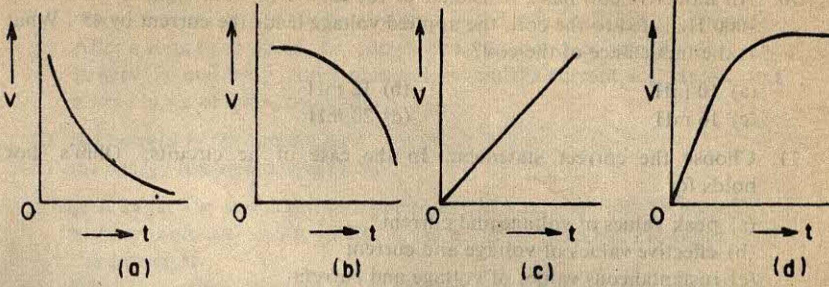


Fig. 8.20

as shown in Fig. 8.21. The key is plugged at $t = 0$. What is the final value of current in the 10 Ω resistor?

(a) $\frac{2}{3}$ A (b) $\frac{1}{3}$ A

(c) $\frac{1}{6}$ A (d) zero

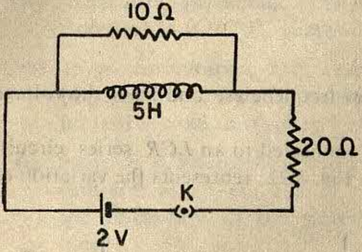


Fig. 8.21

4. What is the final value of the current through the 20 Ω resistor shown in Fig. 8.21?

   (a) zero (b) 0.1 A
   (c) 2/3 A (d) 1/3 A

5. What is the self inductance of a coil in which an induced emf of 2 V is set up when the current changes at the rate 4 A s^{-1}?

   (a) 0.5 mH (b) 0.5 H
   (c) 2.0 H (d) 8.0 H

6. The time constant of a circuit consisting of a capacitance C and a resistor R is

   (a) R/C (b) C/R
   (c) RC (d) $\sqrt{RC}$

7. A 10 Ω electric heater is connected to a 220 V, 50 Hz mains supply. What is the peak value of the potential difference across the heater element?

   (a) 220 V (b) $220/\sqrt{2}$ V
   (c) 110 V (d) $220\sqrt{2}$ V

8. A choke is used as a resistance in

   (a) dc circuits (b) ac circuits
   (c) both ac and dc circuits (d) full-wave rectifier circuits

9. An ac series circuit contains 4 Ω resistance and 3 Ω inductive reactance. What is the impedance of the circuit?

   (a) 1 Ω (b) 5 Ω

   (c) 7 Ω (d) $\frac{7}{\sqrt{2}}\ \Omega$

10. An inductive coil has a resistance of 100 Ω. When an ac signal of frequency 1000 Hz is fed to the coil, the applied voltage leads the current by 45°. What is the inductance of the coil?

(a) 10 mH	(b) 12 mH
(c) 16 mH	(d) 20 mH

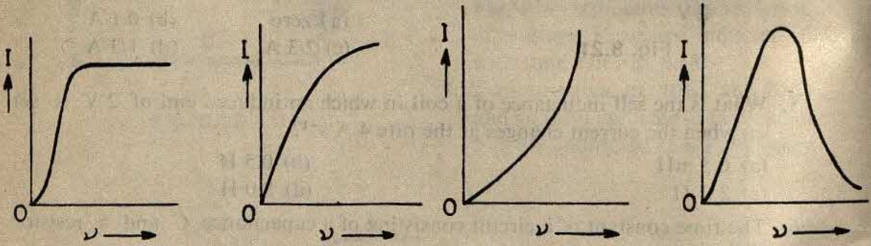
11. Choose the correct statement. In the case of ac circuits, Ohm's law holds for

    (a) peak values of voltage and current
    (b) effective values of voltage and current
    (c) instantaneous values of voltage and current
    (d) all of the above

12. In a purely inductive circuit, the current

    (a) lags behind the voltage by $\pi/2$
    (b) leads the voltage by $\pi/2$
    (c) is in phase with the voltage
    (d) lags behind the voltage by $\pi/2$ at low frequencies, and leads the voltage by $\pi/2$ at high frequencies

13. An ac source of variable frequency ν is connected to an *LCR* series circuit. Which one of the graphs shown in Fig. 8.22 represents the variation of current *I* in the circuit with frequency



(a) (b) (c) (d)

Fig. 8.22

D. Numerical Problems

1. (a) A coil of resistance 5 Ω and inductance 1 H is suddenly connected to a battery (negligible resistance) of 10 V at $t = 0$. Calculate:

    (i) the steady current,
    (ii) the rate of rise of current,
    (iii) the rate at which the energy is supplied by the battery,
    (iv) the rate at which the energy is dissipated as heat, and
    (v) the rate at which the energy is stored in the magnetic field of the coil.

   (b) In problem 1(a), at $t = 0.4$ s, calculate:

    (i) the current in the circuit,
    (ii) the rate of rise of current,
    (iii) the rate at which the energy is supplied by the battery,
    (iv) the rate at which the energy is dissipated as heat, and

(v) the rate at which the energy is stored in the magnetic field of the coil.

(c) After a long time when the current has attained a steady value, the battery is suddenly short-circuited so that the current decreases. After a time lapse of 0.4 s, calculate:

(i) the current in the circuit, and
(ii) the energy released to the circuit.

2. (a) (i) A capacitor is charged to a potential difference of 500 V through a resistor and an energy of 20 J is stored in it. Calculate the capacity of the capacitor.
(ii) What is the time constant of the above *CR* circuit if the capacitor is charged to 80% of its maximum charge in 10 s?
(iii) What is the value of the resistor that gives this time constant.
(iv) What is the percentage of the energy of the charged capacitor that is stored when it is 80% charged.

(b) In a photographic flash, usually a condenser is charged to a high potential and discharged through a resistance to give a momentary flash for 0.01 second, calculate:

(i) the value of resistance which gives a time constant of 0.01 s using the capacitor of part (a), and
(ii) the percentage of the stored energy that is discharged during this time.

3. An *LR* circuit of impedance 400 Ω and a resistance 300 Ω is connected to a (220 V, 50 Hz) supply. Calculate:

(a) the inductance of the circuit, and
(b) the phase angle.

4. A current of 10 A flows in a coil when connected to a 120 V, dc supply and a current of 0.2 A flows when connected to a 120 V ac supply. Calculate:

(a) the resistance of the coil,
(b) the inductance of the coil if the frequency is 60 Hz,
(c) the phase angle, and
(d) the frequency of ac supply to give a phase angle of 45°.

5. A series *RC* circuit ($R = 2\ \text{k}\Omega,\ C = 30\mu\ F$) is connected a (220 V, 50 Hz) supply. Calculate:

(a) the voltage across the capacitor,
(b) the voltage across the resistance,
(c) the phase angle,
(d) the impedance of the circuit,
(e) the instantaneous current as a function of time, and
(f) the current in the circuit.

6. An ac source (220 V, 50 Hz) is connected to a series circuit containing a resistance 200 Ω, an inductance 0.5 H and a capacitance 10 μF. Calculate:

(i) the impedance of the circuit,
(ii) the current in the circuit,
(iii) the voltage across each part, and
(iv) the phase angle.

7. What are the readings of voltmeters V_1, V_2 and V_3 in Fig. 8.23?

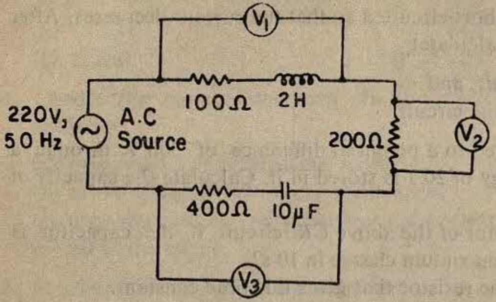


Fig. 8.23

8. An *LCR* series circuit is connected to an ac source of 24 V. At a certain frequency, both the inductive and capacitative reactances are 40 Ω. If $R = 10\ \Omega$, calculate:

   (a) the voltage across R
   (b) the voltage across L,
   (c) the voltage across the series combination of C and L, and
   (d) the frequency of ac source if $L = 0.1$ H.

9

ELECTROMAGNETIC WAVES

9.1 INTRODUCTION

In the preceding chapter we have studied the fundamental facts about electricity and magnetism. They include all but *one* of the basic ideas of the modern theory of electromagnetism. In this chapter we will indicate that one missing link which will complete the description of the basic equations of electromagnetism. In his revolutionary work published in 1865, Maxwell provided that missing link, known as the *displacement current.* This new kind of current (see Sec. 9.2 below) brought together the seemingly unconnected phenomena of electricity and magnetism into a coherent and unified theory. Having discovered the missing link, Maxwell then postulated the transmission of energy by electromagnetic waves. His discovery that electromagnetic waves travel with the speed of light led him to conclude that light itself is an electromagnetic wave, as had been suggested earlier by Faraday.

9.2 MAXWELL'S DISPLACEMENT CURRENT

In Chapter 7 we studied the magnetic fields due to currents. The magnetic field **B** is related to current I by Ampere's circuital law:

$$\oint_C \mathbf{B}\cdot d\mathbf{l} = \mu_0 I \tag{9.1}$$

where I is the current across a closed surface whose periphery is the loop C.

In 1864, Maxwell showed that Eq. (9.1) was logically inconsistent and that something was missing. We give below his reasoning which shows the inconsistency and incompleteness of Eq. (9.1).

Consider a parallel-plate capacitor being charged with a battery (Fig. 9.1). During the charging process, a current I flows through the connecting wires, which of course changes with time, as we have seen in the previous chapter. The current will produce a magnetic field around it which can be detected by a magnetic compass.

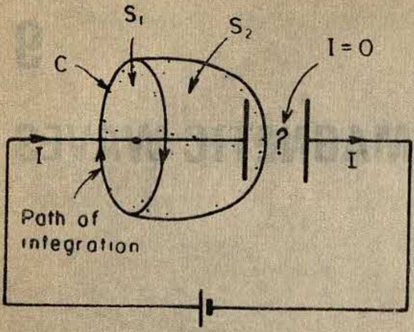


Fig. 9.1 *A parallel-plate capacitor being charged by a battery*

Maxwell's reasoning was as follows: Consider two surfaces S_1 and S_2 bounded by the same closed circular loop C. Surface S_1 is a plane circular surcfae enclosing only the connecting wire attached to the left plate of the capacitor while surface S_2 is a hemispherical surface which lies in the region between the plates. Both surfaces enclose the same loop C, which is the path of the line integral in Eq. (9.1). Now the circular area S_1 bounded by loop C has a current I flowing across it, because the wire passes through it. Hence from Ampere's law, we have

For surface S_1:

$$\oint_C \mathbf{B} \cdot d\mathbf{l} = \mu_0 I \tag{9.2}$$

The surface S_2 encloses one of the capacitor plates and lies in the region between them. No actual current (i.e. current due to the flow of electrons) flows in this region because it contains a dielectric. Since *no current apparently flows out of this surface*, we, have from Ampere's law

For surface S_2:

$$\oint_C \mathbf{B} \cdot d\mathbf{l} = 0 \tag{9.3}$$

The two results (9.2) and (9.3) obviously contradict each other, since the value of the loop integral of magnetic field along the same loop C cannot depend upon the kind of surface bounded by the loop. The two surfaces S_1 and S_2 have an equally good claim to be used.

The two results (9.2) and (9.3) also violate the principle of continuity of current (or conservation of charge), since a current I flows into the surface S_2 but apparently no current flows out it. These considerations led Maxwell to a belief that a current of *some sort* must be flowing in the region between the capacitor plates during the charging process. To remove the contradiction, Maxwell modified Ampere's law [Eq. (9.1)] as follows:

$$\oint \mathbf{B}\cdot d\mathbf{l} = \mu_0 I + \mu_0\epsilon_0 \frac{d\Phi_E}{dt} = \mu_0\left(I + \epsilon_0\frac{d\Phi_E}{dt}\right) \tag{9.4}$$

where Φ_E is the electric flux across the loop. Just as magnetic flux $\Phi_B = BA$ (magnetic field × area), the electric flux is $\Phi_E = EA$ (electric field × area). Notice that $\epsilon_0\, d\Phi_E/dt$ has the dimensions of current. Maxwell named this product as the displacement current I_d, i.e.

$$\text{Displacement current } I_d = \epsilon_0 \frac{d\Phi_E}{dt} \tag{9.5}$$

so that the generalized Ampere's law reads:

$$\oint \mathbf{B}\cdot d\mathbf{l} = \mu_0(I + I_d) \tag{9.6}$$

where I is the usual conduction current (i.e. current caused by the conduction of electrons). Unlike the conduction current I, the displacement current I_d exists as long as the electric field E (and hence the electric flux Φ_E) is changing with time. Maxwell's important contribution is that a *time-varying electric field produces a magnetic field* just as a conduction current does.

Continuity of $(I + I_d)$

The sum $(I + I_d)$ has the important property of continuity along any closed path, although individually they may not be continuous. To show this, let us again consider Fig. (9.1). A current I enters one plate and leaves the other plate of the capacitor. The *conduction* current I is *not continuous* across the capacitor gap because no charge is transported across this gap. If the sum $(I + I_d)$ is to be continuous, the displacement current I_d in the gap must be equal to the conduction current I in the lead wires. This can be shown as follows.

We know that the electric field E in the gap of a parallel plate capacitor of plate area A is given by

$$E = \frac{q}{\epsilon_0 A}$$

where q is the charge on the capacitor plates at any instant of time during the charging process. Differentiating, we get

$$\frac{dE}{dt} = \frac{1}{\epsilon_0 A}\frac{dq}{dt} = \frac{I}{\epsilon_0 A} \tag{9.7}$$

By definition, the displacement current I_d is

$$\begin{aligned} I_d &= \epsilon_0 \frac{d\Phi_E}{dt} \\ &= \epsilon_0 \frac{d}{dt}(EA) \\ &= \epsilon_0 A \frac{dE}{dt} \qquad (9.8) \end{aligned}$$

Combining Eqs (9.7) and (9.8) we get

$$I_d = \epsilon_0 A \frac{I}{\epsilon_0 A} = I$$

which is exactly the value of conduction current in the lead wires. The sum $(I + I_d)$ thus has the same value along the entire path, although individually they are discontinuous.

Note: We know from Faraday's law of induction, that

$$\oint \mathbf{E} \cdot d\mathbf{l} = -\frac{d\Phi_B}{dt} \qquad (9.9)$$

where Φ_B is the magnetic flux. This law asserts that an electric field [LHS of (9.9)] is produced by a changing magnetic flux (or magnetic field) [RHS of (9.9)]. Does the counterpart of Eq. (9.9) exist? According to Maxwell, it does and from considerations of symmetry, the counterpart is

$$\oint \mathbf{B} \cdot d\mathbf{l} = \mu_0 \epsilon_0 \frac{d\Phi_E}{dt} = \mu_0 I_d \qquad (9.10)$$

where $(\mu_0 \epsilon_0)$ has to be introduced to keep the dimensions of both sides consistent. Equation (9.10) asserts that a magnetic field [LHS of (9.10)] can be produced by a changing electric field [RHS of (9.10)] In each case, the appropriate flux Φ_B or Φ_E is *increasing* in Eq. (9.9) or (9.10). However, experiment shows that the field lines of **E** are *counter-clockwise* while those of **B** are *clockwise.* This difference requires that the minus sign in Eq. (9.9) be omitted from Eq. (9.10). Hence the analogy between the two is complete.

We also know that a magnetic field is also produced by a current, i.e. $\oint \mathbf{B} \cdot d\mathbf{l} = \mu_0 I$. Thus there are at least two ways of producing a magnetic field: (a) by a time-varying electric field and (b) by a current. Maxwell said that both these possibilities must be allowed and therefore, Ampere's law be modified as

$$\oint \mathbf{B} \cdot d\mathbf{l} = \mu_0 (I + I_d) = \mu_0 \left(I + \epsilon_0 \frac{d\Phi_E}{dt}\right)$$

In situation represented by Eq. (9.2) no time-varying electric field is present, hence the term $d\Phi_E/dt$ is zero and in the situation represented by Eq. (9.3) there is no conduction current in the gap, so the term I is zero but the RHS of Eq. (9.3) should be $\mu_0\epsilon_0\, d\Phi_E/dt$, not zero, since the time-varying flux is not zero.

EXAMPLE 9.1 Figure 9.2 shows a capacitor made of two circular plates, each of radius $R = 12$ cm, separated by $d = 5.0$ mm. The capacitor is being charged by an external source (not shown in the figure). The charging current I is constant and equal to 0.15 A.

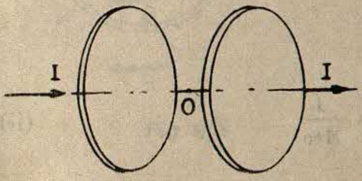


Fig. 9.2

(a) Calculate the capacitance and the rate of change of potential difference between the plates.

(b) Obtain the displacement current across the plates.

(c) Is Kirchhoff's first rule valid at each plate of the capacitor? explain.

Solution:

$$\text{Radius} \quad (R) = 12 \text{ cm} = 12\times 10^{-2} \text{ m}$$

$$\text{Plate area } (A) = \pi R^2 = \pi\times(12\times 10^{-2})^2 \text{ m}^2$$

$$\text{Plate separation} \quad (d) = 5.0 \text{ mm} = 5.0\times 10^{-3} \text{ m}$$

(a) The capacitance is given by

$$C = \frac{\epsilon_0 A}{d}$$

$$= \frac{1}{4\pi(9\times 10^9)}\times\frac{\pi(12\times 10^{-2})^2}{5.0\times 10^{-3}} = 80.1\times 10^{-12} \text{ F}$$

Now instantaneous charge q on capacitor plates is

$$q = CV$$

$$\therefore \quad I = \frac{dq}{dt} = C\frac{dV}{dt}$$

$$\therefore \quad \frac{dV}{dt} = \frac{I}{C} = \frac{0.15}{80.1\times 10^{-12}} = 1.87\times 10^9 \text{ Vs}^{-1}$$

(b) The displacement current is

$$I_d = \epsilon_0 \frac{d\Phi_E}{dt}$$

$$= \epsilon_0 \frac{d}{dt}(EA)$$

$$= \epsilon_0 A \frac{dE}{dt} \qquad \text{(i)}$$

We know that $E = \dfrac{q}{A\epsilon_0}$

$$\therefore \quad \frac{dE}{dt} = \frac{1}{A\epsilon_0}\frac{dq}{dt} = \frac{I}{A\epsilon_0} \qquad \text{(ii)}$$

Using (ii) in (i), we have

$$I_d = \epsilon_0 A \frac{I}{\epsilon_0 A} = I = 0.15 \text{ A}$$

This result also follows from the fact that $(I + I_d)$ is continuous.

(c) Since the sum $(I + I_d)$ is continuous, Kirchhoff's law for currents holds if the term "current" in the law includes the displacement current as well.

Example 9.2 Referring to Fig. (9.2), calculate: (a) the magnetic field at a point (i) on the axis, (ii) 6.5 cm from the axis and (iii) 15 cm from the axis. (b) At what distance from the axis is the magnetic field due to the displacement current the greatest? Find the maximum value of the field. Use the data given in Example 9.1.

Solution: The magnetic field is given by (since there is no conduction current in the gap of the capacitor)

$$\oint_C \mathbf{B}\cdot d\mathbf{l} = \mu_0(I + I_d)$$

$$= \text{current through the area of loop C} \qquad \text{(i)}$$

We will now evaluate the loop integral in Eq. (i) at a point at a distance r from the axis for the two cases: $r \leqslant R$ and $r \geqslant R$, where R is the radius of each plate of the capacitor.

Case: $r < R$ Consider a circle of radius r ($< R$) between the circular plates and coaxial with them as shown in Fig. 9.3. The common centre lies on the axis of the plates and plane of the circle is perpendicular to the axis (which is perpendicular to the page). The magnetic field **B** has the same magnitude at every point on the circle and its direction is tangential at every point as shown. Therefore, the line integral along the circular loop is

$$\oint \mathbf{B}\cdot d\mathbf{l} = \mathbf{B}\int_{\text{circle}} dl = B(2\pi r)$$

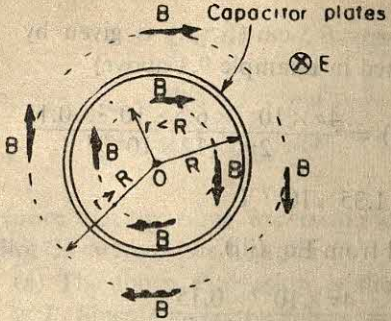


Fig. 9.3

From Ampere's law, Eq. (i) we have (since conduction current $I = 0$ in the region between the plates),

$B(2\pi r)$ = displacement current passing through the area enclosed by the circle of radius r

= $\mu_0\epsilon_0$ (rate of change of electric flux through the area πr^2)

Now let Φ_E be the flux through the plates of area πr^2. Since the flux $\propto$ area, the flux through the area πr^2 (which is less than πR^2) is $\Phi_E \dfrac{\pi r^2}{\pi R^2} = \Phi_E \dfrac{r^2}{R^2}$. Therefore, the rate of change of electric flux through the area $\pi r^2 = \left(\dfrac{d\Phi_E}{dt}\right)\cdot\dfrac{r^2}{R^2}$. Hence

$$B(2\pi r) = \mu_0\epsilon_0 \frac{d\Phi_E}{dt}\cdot\frac{r^2}{R^2}$$

But $\epsilon_0\, d\Phi_E/dt = I_d$, the displacement current through the capacitor plates. Therefore, we have

$$B(2\pi r) = \mu_0 I_d \frac{r^2}{R^2}$$

or

$$B = \frac{\mu_0 r I_d}{2\pi R^2} \qquad (\text{for } r \leqslant R) \qquad \text{(ii)}$$

Case: $r \geqslant R$ Consider now the circle of radius $r \geqslant R$ as shown in Fig. (9.3). Since the area πr^2 of this larger circle now completely encloses the capacitor plates, it is obvious that the total current passing through πr^2 is simply I_d.

Hence for, $r \geqslant R$,

$$B(2\pi r) = \mu_0 I_d$$

or

$$B = \frac{\mu_0 I_d}{2\pi r} \qquad (\text{for } r \geqslant R) \qquad \text{(iii)}$$

(i) From Eq. (ii) it follows that the magnetic field on the axis of the plates (i.e. for $r = 0$) $= 0$

(ii) Now $R = 12$ cm. The field at $r = 6.5$ cm ($< R$) is given by Eq. (ii) where $I_d = 0.15$ A (as calculated in Example 9.1 above)

$$\therefore \quad B \text{ at } r(= 6.5\,\text{cm} = 6.5 \times 10^{-2}\,\text{m}) = \frac{4\pi \times 10^{-7} \times 6.5 \times 10^{-2} \times 0.15}{2\pi \times (12 \times 10^{-2})^2}$$

$$= 1.35 \times 10^{-7}\,\text{T}$$

(iii) Field at $r = 15$ cm is obtained from Eq. (iii).

$$B \text{ at } r\,(= 15 \times 10^{-2}\,\text{m}) = \frac{4\pi \times 10^{-7} \times 0.15}{2\pi \times (15 \times 10^{-2})}$$

$$= 2.0 \times 10^{-7}\,\text{T}$$

(b) It is clear that B is maximum at $r = R$ in which case both Eqs (ii) and (iii) reduce to

$$B_{\max} = \frac{\mu_0 I_d}{2\pi R} = \frac{4\pi \times 10^{-7} \times 0.15}{2\pi \times (12 \times 10^{-2})^2} = 2.5 \times 10^{-7}\,\text{T}$$

EXAMPLE 9.3 (a) Use Ampere's law to determine the magnetic field due to the conduction current outside the plates (refer to Fig. 9.2 of Example 9.1 and use the data given there) at points: (i) 6.5 cm, (ii) 12 cm, and (iii) 15 cm from the wire. Compare your answers with those in Example 9.2 above.

(b) If the conducting wire has a radius of 1.0 mm, what is the maximum value of the magnetic field due to the conduction current. Compare your result with the answer obtained in Example 9.2(b).

Solution: Figure 9.4 shows the cross-sectional view of the wire placed perpendicular to the plane of the page and a current I directed into the page. The magnetic field lines are concentric circles in the plane of the page. The magnetic field has the same constant value at every point on the circumference of the circle drawn at a distance r from wire; the direction of **B** being along the tangent to the circle at every point. Hence

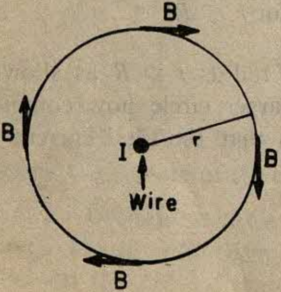


Fig. 9.4

$$\oint \mathbf{B} \cdot d\mathbf{l} = B(2\pi r).$$

From Ampere's law:

$$\oint \mathbf{B}\cdot d\mathbf{l} = \mu_0 \times \text{current through the circle}$$

or $$B(2\pi r) = \mu_0 I$$

or $$B = \frac{\mu_0 I}{2\pi r}$$

Notice that the same expression can also be obtained by using the Biot-Savart law (see Chapter 5).

(a) The above expression is the same as that for B for the case $r \geqslant R$. Since $I = I_d$, the values of B for $r = 12$ cm and $r = 15$ cm are the same as in Example 9.2 above. But for $r = 6.5$ cm, the expressions are different. In Example 9.2, it was 1.35×10^{-7} T.

(b) Since $B \propto 1/r$, the value of B is maximum at the surface of the wire, i.e. at $r = 1.0$ mm $= 1.0\times 10^{-3}$ m.

$$B_{\max} = \frac{4\pi\times 10^{-7}\times 0.15}{2\pi\times(1.0\times 10^{-3})}$$

$$= 3.0\times 10^{-5}\ \text{T}$$

which is much greater than the corresponding value (2.5×10^{-7} T) obtain in Example 9.2(b). The magnetic fields due to conduction currents can be multiplied considerably by using a coil of many turns, and the magnetic fields due to displacement currents can be increased in experiments involving oscillations at very high frequencies so that dE/dt becomes very high.

EXAMPLE 9.4 The capacitor shown in Fig. 9.2 of Example 9.1 has a capacitance of 100 pF and is connected to a 220 V, 50 Hz ac power supply. (a) Find the rms value of the conduction current. (b) Is the conduction current equal to the displacement current? (c) Find the amplitude of **B** at a point 6.0 cm from the axis between the plates. Use the data given in Example 9.1.

Solution:

$$C = 100\ \text{pF} = 10^{-10}\ \text{F}$$

$$V_{\text{rms}} = 220\ \text{V}$$

$$\omega = 2\pi\times 50 = 100\ \pi\ \text{rad s}^{-1}$$

(a) Impedance of capacitor is

$$Z_C = \frac{1}{\omega C}$$

Now $$I_{\rm rms} = \frac{V_{\rm rms}}{Z_C} = V_{\rm rms}\cdot\omega C$$

$$= 220\times 100\pi\times 10^{-10} = 6.9\times 10^{-6}\ \text{A}$$

$$= 6.9\ \mu\text{A}$$

(b) Yes, the derivation given in Example 9.1(b) is true for oscillating electric fields, i.e. $I = I_d$.

(c) The instantaneous value of B is

$$B = \frac{\mu_0 r I}{2\pi R^2}$$

The amplitude of B is ($\because$ B and I are in phase)

$$B_0 = \frac{\mu_0 r I_0}{2\pi R^2}$$

where $$\text{Amplitude of } I = I_0 = \sqrt{2}\, I_{\rm rms}$$

$$= \sqrt{2}\times 6.9\ \mu\text{A}$$

$$= 9.76\ \mu\text{A}$$

9.3 ELECTROMAGNETIC WAVES

Consider a region where there are no charges and conduction currents ($I = 0$). Such a region is called a dielectric, including vacuum. Here we consider the simplest case of waves in vacuum. For vacuum. Ampere's law [Eq. (9.4)] and Faraday's law [Eq. (9.5)] take the symmetrical form.

$$\oint \mathbf{B}\cdot d\mathbf{l} = \mu_0\epsilon_0 \frac{d\Phi_E}{dt} \tag{9.11}$$

$$\oint \mathbf{E}\cdot d\mathbf{l} = -\frac{d\Phi_B}{dt} \tag{9.12}$$

The former states that a time-varying electric field at any point is a source of magnetic field and the latter states just the reverse, namely, that a time-varying magnetic field is a source of electric field. It is reasonable to think that a wave of electric and magnetic fields, both varying with space and time, will exist; one providing the source of the other. It can be rigorously proved from Eqs (9.11) and (9.12) that such a wave (called the *electromagnetic wave*) does indeed exist. We will give a somewhat indirect proof. We will assume an electromagnetic wave and show that Eqs (9.11) and (9.12) can be satisfied. We will then calculate the speed of such a wave.

Consider a plane electromagnetic wave travelling in the x-direction. We will show (a little later in this section) that, for a plane wave, the electric and magnetic fields are transverse to the direction of travel (i.e. they are in the y-z plane). For the moment we assume this. We further assume that the electric field is along the y-direction. This field E_y depends only on x and t (for a plane wave). The magnetic field will then be along the z-direction, perpendicular to E_y. The field B_z also depends only on x and t. Figure 9.5 shows such a plane wave travelling in the x direction. The **E**-field is along the y-axis and the **B**-field is along the z-axis.

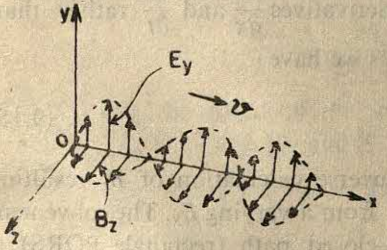


Fig. 9.5 *Schematic representation of a plane electromagnetic wave travelling along the x-axis*

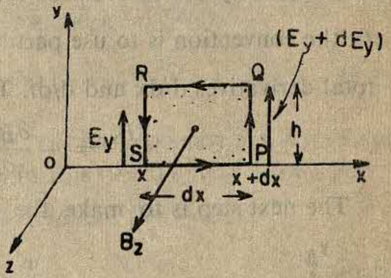


Fig. 9.6 *Generation of B_z by a varying E_y*

Figure 9.6 is used to show the application of Eq, (9.12) to the plane E wave moving in the x-direction. A convenient closed path (rectangle PQRS) is chosen in the x-y plane, around which we shall take the line integral of **E**. This is equated through Eq. (9.12) to the rate of change of flux of $\mathbf{B}(\Phi_B)$ through the plane bounded by the path of the line integral.

Let E_y be the E-field at x and $(E_y + \mathrm{d}E_y)$ that at $(x + \mathrm{d}x)$, here $\mathrm{d}x$ is length of the rectangle. Let its height be h. It is clear that the line integral of **E** along the horizontal parts PQ and RS is zero because these paths are along the direction of the E_y fields so that $(\mathbf{E}\cdot\mathrm{d}\mathbf{l}) = 0$ along the y-direction. Only the vertical parts PS and QR contribute. The line integral then becomes

$$\oint \mathbf{E}\cdot\mathrm{d}\mathbf{l} = [(E_y + \mathrm{d}E_y)h - (E_y h)]$$

$$= \mathrm{d}E_y \cdot h \tag{9.13}$$

The magnetic flux Φ_B for the rectangle is

$$\Phi_B = B_z \times \text{area} = B_z h\,\mathrm{d}x$$

Differentiating this, we get

$$\frac{\mathrm{d}\Phi_B}{\mathrm{d}t} = (h\,\mathrm{d}x)\frac{\mathrm{d}B_z}{\mathrm{d}t} \tag{9.14}$$

Using Eqs (9.13) and (9.14) in Eq. (9.12) we get

$$\mathrm{d}E_y \cdot h = -(h\,\mathrm{d}x)\,\frac{\mathrm{d}B_z}{\mathrm{d}t}$$

or

$$\frac{\mathrm{d}E_y}{\mathrm{d}x} = -\frac{\mathrm{d}B_z}{\mathrm{d}t}$$

Since both E_y and B_z are functions of more that one variable (x and t), the convention is to use partial derivatives $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial t}$ rather than total derivatives $\mathrm{d}/\mathrm{d}x$ and $\mathrm{d}/\mathrm{d}t$. Thus we have

$$\frac{\partial E_y}{\partial x} = -\frac{\partial B_z}{\partial t} \tag{9.15}$$

The next step is to make the converse calculation of B_z resulting from a varying E_y. The convenient closed path (rectangle PQRS) is chosen in the x-z plane, around which we will take the line integral of **B**, This is equated through Eq. (9.11) to $\mu_0\epsilon_0$ times the rate of change of electric flux of $\mathbf{E}(\Phi_E)$ through the plane bounded by the path of the line integral. We find, by a similar calculation, that

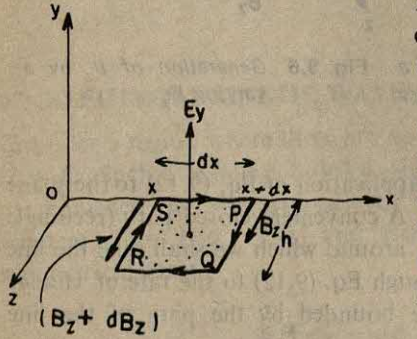


Fig. 9.7 *Generation of E_y by a varying B_z*

$$\oint \mathbf{B}\cdot \mathrm{d}\mathbf{l} = [(B_z h) - (B_z + \mathrm{d}B_z)(h)]$$

$$= -\mathrm{d}B_z \cdot h \tag{9.16}$$

The electric flux through the rectangle is

$$\Phi_E = E_y \times \text{area} = E_y(h\,\mathrm{d}x)$$

Differentiating we have

$$\frac{\partial \Phi_E}{\partial t} = (h\,\mathrm{d}x)\frac{\partial E_y}{\partial t} \tag{9.17}$$

Using Eqs (9.16) and (9.17) in Eq. (9.11), we get

$$-(\mathrm{d}B_z)h = \mu_0\epsilon_0\,(\mathrm{d}x\,h)\frac{\partial E_y}{\partial t}$$

or

$$\frac{\partial B_z}{\partial x} = -\mu_0\epsilon_0 \frac{\partial E_y}{\partial t} \tag{9.18}$$

Now we assumed that the electromagnetic wave is travelling in the x-direction. The E_y and B_z fields are, therefore, given by (see Fig. 9.5)

$$E_y = E_0 \sin(\omega t - kx) \tag{9.19}$$

$$B_z = B_0 \sin(\omega t - kx) \tag{9.20}$$

where E_0 and B_0 are the amplitudes of E_y and B_z respectively, ω is the angular frequency and k is the wave number of the wave. Quantities ω and k are related as (since $\omega = 2\pi\nu$ and $k = 2\pi/\lambda$):

$$\frac{\omega}{k} = \frac{2\pi\nu}{2\pi/\lambda} = \nu\lambda = v \tag{9.21}$$

where v is the speed of the wave, ν is its frequency and λ its wavelength. Using Eqs (9.19) and (9.20) in Eqs (9.15) and (9.18) we get

$$-kE_0 \cos(\omega t - kx) = \omega B_0 \cos(\omega t - kx)$$

or

$$\frac{E_0}{B_0} = -\frac{\omega}{k} \tag{9.22}$$

and

$$-kB_0 \cos(\omega t - kx) = \mu_0\epsilon_0 E_0 \cos(\omega t - kx)$$

or

$$\frac{B_0}{E_0} = -\frac{\mu_0\epsilon_0\omega}{k} \tag{9.23}$$

Eliminating E_0 and B_0 from Eqs (9.22) and (9.23) we have

$$\frac{\omega^2}{k^2} = \frac{1}{\mu_0\epsilon_0}$$

which with the use of Eq. (9.21) gives

$$v = \frac{1}{\sqrt{\mu_0\epsilon_0}} \tag{9.24}$$

which is the speed of electromagnetic waves in vacuum. Now, for vacuum, we know that $\mu_0 = 4\pi \times 10^{-7}$ TmA^{-1} and $\epsilon_0 = 8.85 \times 10^{-12}$ C^2 N^{-1} m^{-2}. Substituting these values in Eq. (9.24) we have

$$v = \frac{1}{[(4\pi \times 10^{-7})(8.85 \times 10^{-12})]^{1/2}} = 3.0 \times 10^8 \text{ ms}^{-1}$$

which is the speed of light in vacuum measured experimentally. The excellent agreement between the experimentally measured speed of light ($c = 2.997924 \times 10^8$ ms^{-1}) to such a high degree of accuracy and the value based on the experimental measurements of ϵ_0 and μ_0

gave the first quantitative proof of the fact that light is an electromagnetic wave. The emergence of the speed of light from purely electromagnetic considerations is the crowning achievement of Maxwell's electromagnetic theory.

Transverse Nature of Plane Electromagnetic Waves

We now show that for a plane electromagnetic wave travelling, say, in the x direction, the directions of the **E** and **B** fields are transverse to the x-direction, the direction of propagation, i.e. **E** and **B** are confined to the yz plane. To show this we will use Gauss's law in electricity and its analogue in magnetism (see Chapter 7); they are

$$\oint_S \mathbf{E}\cdot d\mathbf{S} = \frac{q}{\epsilon_0}$$

and

$$\oint_S \mathbf{B}\cdot d\mathbf{S} = 0 \tag{9.25}$$

since there are no free charges in vacuum ($q = 0$), the former reduces to

$$\oint_S \mathbf{E}\cdot d\mathbf{S} = 0 \tag{9.26}$$

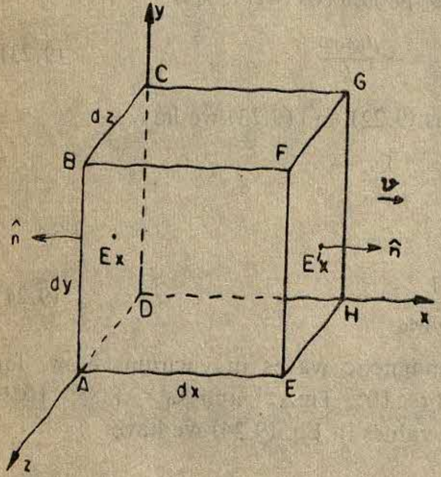


Fig. 9.8 *Elementary cube ABCDEFGH used to prove that a plane electromagnetic wave is transverse*

To perform the surface integrals in Eqs (9.25) and (9.26) we take an elementary cube ABCDEFGH of sides dx, dy and dz (Fig. 9.8). First we apply Eq. (9.26) to the cube. Since we are dealing with plane waves, E does not vary with y or z, so the flux of E through faces ADEH, BCGF, ABEF and CDHG is zero. We are left with faces ABCD and EFGH. Let E_x be the x-component of **E** at face ABCD and E'_x that at face EFGH. The flux across ABCD is

$$\int_{\mathrm{ABCD}} \mathbf{E}\cdot d\mathbf{S} = -E_x\,(dy\,dz)$$

since E_x is normal to this face and the face area is $dy\,dz$ (see Fig. 9.8). The minus sign results from the negative direction (along-x axis) of the outward normal to this face. Similarly the flux across EFGH is

$$\int_{\mathrm{EFGH}} \mathbf{E}\cdot d\mathbf{S} = +E'_x\,(dy\,dz)$$

Since the total flux through the whole cube is zero, we must have

$$-E_x(dy\,dz) + E'_x(dy\,dz) = 0$$

or

$$E'_x = E_x$$

i.e. E_x must have the same value at both the faces at all times. Therefore E_x cannot vary with x, i.e. E_x = constant. Now, a constant field cannot give rise to a wave, only a space-varying field can. Hence, if we are interested only in waves, E_x can be set equal to zero. Similarly, using Eq. (9.25) we can show that B_z = constant and has nothing to do with wave motion. We thus see that, *in a plane electromagnetic wave, the electric and magnetic fields are transverse*, i.e. perpendicular to the direction of wave propagation. Thus, for a plane wave propagation in the x-direction, the **E** and **B** fields are in the yz plane. In our example we have considered one combination (E_y, B_z). The other combination (E_z, B_y) also exists. When both pairs are present we have, what is called, an *unpolarized wave*. If only one pair, (E_y, B_z) or (E_z, B_y) is present, the wave is said to be *polarized*, as we will see in the next chapter.

Production of Electromagnetic Waves

Consider a charge at rest. It has an electric field in the region around it but no magnetic field. If it is given an impulse so that it begins to move, it produces electric and magnetic fields. If it is moving with a constant velocity (i.e. if the current is not changing with time) the magnetic field will not change with time, so it cannot produce an electromagnetic wave. But if the charge is somehow accelerated, the magnetic and electric fields will change with space and time; it then produces an electromagnetic wave. Thus *an accelerated charge emits an electromagnetic wave.*

In chapter 8 (Sec. 8.6) we discussed oscillations of an *LC* circuit. We found the charge oscillating across the capacitor plates. An oscillating charge has a non-zero acceleration; hence it will emit an electromagnetic wave which has the same frequency as the oscillating charge has.

In Chapter 13 we will learn that an electron circulating round its nucleus in a stable orbit does not emit an electromagnetic wave, although it is accelerating; it does so only when it falls to a lower orbit.

When fast-moving electrons hit a metal target, electromagnetic waves (x-rays) are produced. The detailed discussion about the production of electromagnetic waves is beyond the scope of this book.

Energy Densities of Electric and Magnetic Fields

Just as an oscillating pendulum has energy associated with it, oscillating electric and magnetic fields also have electric and magnetic energies associated with them. The energy density (i.e. energy per unit volume) of an electric field **E** is given by

$$u_E = \frac{1}{2}\epsilon_0 \,|\,\mathbf{E}\,|^2 = \frac{\epsilon_0}{2}(E_x^2 + E_y^2 + E_z^2) \qquad (9.27)$$

The energy density of a magnetic field is given by

$$u_B = \frac{1}{2\mu_0}\,|\,\mathbf{B}\,|^2 = \frac{1}{2\mu_0}(B_x^2 + B_y^2 + B_z^2) \qquad (9.28)$$

Consider a plane electromagnetic wave travelling in the x-direction with the **E** field confined in the y direction and the **B** field in the z-direction. We can then write

$$E_x = 0 \qquad E_z = 0,$$

$$E_y = E_0 \sin(\omega t - kx)$$

Now

$$\omega = 2\pi\nu = 2\pi/T \text{ and } k = 2\pi/\lambda$$

Therefore,

$$E_y = E_0 \sin\left\{\frac{2\pi}{T}\left(t - \frac{x}{v}\right)\right\} \qquad (9.29)$$

Also $B_x = 0$, $B_y = 0$, and

$$B_z = B_0 \sin\left\{\frac{2\pi}{T}\left(t - \frac{x}{v}\right)\right\} \qquad (9.30)$$

where E_0 and B_0 are the amplitudes of E_y and B_z fields respectively.

For such a wave,

$$u_E = \frac{\epsilon_0}{2} \cdot E_y^2 \tag{9.31}$$

and

$$u_B = \frac{1}{2\mu_0} \cdot B_z^2 \tag{9.32}$$

Using Eq. (9.29) in Eq. (9.31), the average energy density of the electric field of the wave is

$$\langle u_E \rangle = \frac{\epsilon_0}{2} \frac{E_0^2}{T} \int_0^T \sin^2 \left\{ \frac{2\pi}{T}\left(t - \frac{x}{v}\right) \right\} \mathrm{d}t \tag{9.33}$$

The function $\sin^2 \left\{ \frac{2\pi}{T}\left(t - \frac{x}{v}\right) \right\}$ at any x varies with time between the limiting values of 0 and 1 so that the average value of this function is 1/2 and the integral in Eq. (9.33) is just $T/2$. Hence

$$\langle u_E \rangle = \frac{\epsilon_0}{4} \cdot E_0^2 \tag{9.34}$$

Similarly, using Eq. (9.30) in Eq. (9.32), the average energy density of the magnetic field of the wave is

$$\langle u_B \rangle = \frac{1}{4\mu_0} \cdot B_0^2 \tag{9.35}$$

Now from Eq. (9.22) we have (since $\omega/k = v = 1/\sqrt{\mu_0 \epsilon_0}$)

$$B_0^2 = \frac{E_0^2}{v} = \mu_0 \epsilon_0 E_0^2 \tag{9.36}$$

Using Eq. (9.36) in Eq. (9.35), we find that $\langle u_B \rangle = \frac{\epsilon_0}{4} \cdot E_0^2$, the same as Eq. (9.34). Thus we conclude that in a plane electromagnetic wave, the average energy densities of electric and magnetic fields are equal. The total average energy density of the electromagnetic field is

$$u = \langle u_E \rangle + \langle u_B \rangle = \frac{\epsilon_0}{2} E_0^2 = \frac{B_0^2}{2\mu_0} \tag{9.37}$$

EXAMPLE 9.5 In a plane electromagnetic wave, the electric field oscillates sinusoidally at a frequency of 2.0×10^{10} Hz and amplitude 48 $\mathrm{Vm^{-1}}$. (a) What is the wavelength of the wave? (b) What is the amplitude of the oscillating magnetic field? (c) Find the total average energy density of the electromagnetic field of the wave.

Solution:

$$E_0 = 48 \text{ Vm}^{-1} \qquad \nu = 2.0 \times 10^{10} \text{ Hz}$$

Speed of the wave is

$$v = 3 \times 10^8 \text{ ms}^{-1}$$

(a) Wavelength $\lambda = \frac{v}{\nu} = \frac{3.0 \times 10^8}{2.0 \times 10^{10}} = 1.5 \times 10^{-2}$ m

(b) From Eq. (9.22),

$$|B_0| = \frac{k}{\omega} |E_0| = \frac{E_0}{v}$$

$$= \frac{48}{3.0 \times 10^8} = 1.6 \times 10^{-7} \text{ T}$$

(c) From Eq. (9.37),

$$u = \frac{\epsilon_0}{2} E_0^2 = \frac{8.85 \times 10^{-12} \times (48)^2}{2}$$

$$= 1.0 \times 10^{-8} \text{ Jm}^{-3}$$

9.4 EXPERIMENTAL CONFIRMATION OF ELECTROMAGNETIC WAVES

In 1865, Maxwell predicted the existence of electromagnetic waves. His theory further predicted that electromagnetic waves of all frequencies (and hence all wavelengths) should propagate with the speed of light. This theory was first experimentally verified by a German physicist, Heinrich Hertz in 1888. He used a simple electric oscillator and was able to pick up its radiation of electromagnetic waves on a radio receiver some distance away. In 1899, an Italian engineer Guglielmo Marconi succeeded in transmitting electromagnetic waves across the English Channel and 1901 across the Atlantic Ocean. Our radio, radar and TV today are merely elaborations and modifications of those early beginnings; their transmission and reception still depend on oscillating circuits basically similar to those used by Hertz.

HERTZ'S EXPERIMENT

On the basis of theoretical considerations Maxwell asserted that light is a form of electromagnetic radiation. It was about twenty years later that Hertz produced electromagnetic waves of much lower frequencies than those of visible light, detected them and established experimentally that they differed from light only in respect of frequency (or wavelength). He produced electromagnetic waves by using the oscillatory *LC* circuit [see Chapter 8]. The oscillatory current in the circuit produces electromagnetic waves which travel in the space surrounding the *LC* circuit. To ensure that the electromagnetic disturbance is easily detectable, the region of influence had to be extended. This was achieved by using large capacitor plates and separating them by a large distance. The oscillatory current obtained by the discharge of a capacitor through an inductance decays in amplitude and a single oscillatory discharge is of too short a duration to be readily detected. Hertz solved this problem by using a spark gap method. In modern techniques this difficulty is overcome by using the amplifying property of a transistor is an amplifier.

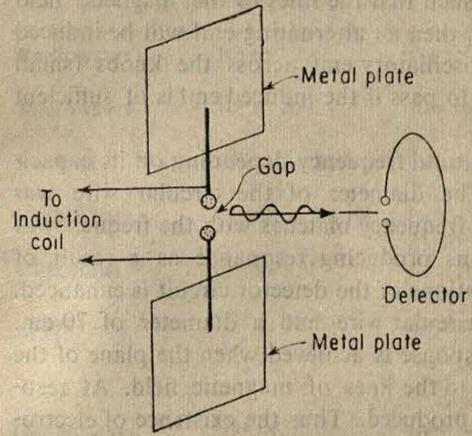


Fig. 9.9 *Schematic illustration of Hertz's experiment*

Hertz's experiment is schematically illustrated in Fig. 9.9. He took two square plates of brass with sides of length 40 cm placed vertically with their centres about 60 cm apart. The opposing faces of these plates were connected to thick wires carrying highly polished brass spheres whose centres were separated by a gap of 2 to 3 cm. To obtain adequate power for the supply to the circuit, Hertz used an induction coil which was connected across the wires attached to the plates, thus charging the capacitor.

When the uni-directional voltage from the induction coil becomes high, the air in the gap between the spheres gets ionized, i.e. the atoms of the gases in air are stripped of their electrons. Consequently, the air in the gap becomes conducting and provides a path for the discharge between the plates. The oscillatory discharge through the conducting air produces a train of electromagnetic waves. Eventually the air in the gap ceases to conduct because the voltage across the capacitor plates falls to too low a value. However, the process is repeated every time the primary circuit of the induction coil is broken. Thus a series of trains of electromagnetic waves is produced.

The period of oscillation of Hertz's circuit was about $T=1.8\times 10^{-8}$ s. Putting the speed of electromagnetic waves in air as $c = 3\times 10^{8}$ ms^{-1}, the wavelength of the waves was $\lambda=cT=(3\times 10^{8}\text{ ms}^{-1})\times(1.8\times 10^{-8}\text{ s})$ $= 5.4$ m and the frequency of the waves was $\nu = 1/I = 1/1.8\times 10^{-8}$ s $= 55$ MHz. Such waves are called radiowaves.

To detect these waves, Hertz used an almost closed circular stout wire terminating at each end by a small polished brass sphere (Fig. 9.9). If this circular wire is placed such that the lines of the magnetic field of the waves pass through it, then an alternating emf will be induced round the circular wire. The oscillatory emf across the knobs (small spheres) causes a small spark to pass if the induced emf is of sufficient magnitude.

The detector has its own natural frequency depending on its capacitance and self-inductance. The diameter of the circular wire was adjusted so that its natural frequency matches with the frequency of the electromagnetic waves, thus producing resonance as a result of which the amplitude of oscillations of the detector circuit is enhanced. Hertz used a detector whose circular wire had a diameter of 70 cm. The maximum resonator response is achieved when the plane of the circular wire is perpendicular to the lines of magnetic field. At resonance, the spark is readily produced. Thus the existence of electromagnetic waves around an oscillatory circuit was detected. Seven years later, an Indian physicist Jagdish Chandra Bose succeeded in producing and observing electromagnetic waves of much shorter wavelength ($\sim$10 mm). These waves are called microwaves.

Hertz also demonstrated experimentally that electromagnetic waves can be reflected and refracted like visible light. He also showed that these waves exhibit interference phenomena just like visible light does, thus establishing beyond doubt that the radiation is in the form of waves.

9.5 SPECTRUM OF ELECTROMAGNETIC RADIATION

Beginning with the remarkable demonstration by Hertz of the existence of long wavelength electromagnetic waves, scientists began looking for electromagnetic waves of wavelengths much shorter than the visible light. In 1898 Röntgen discovered x-rays which are electromagnetic waves of wavelength about 10^{-10} m. The wavelength of visible light is the range of 4×10^{-7} to 8×10^{-7} m. Radiowaves, x-rays and visible light are all electromagnetic waves and travel with the same speed $c = 3.0 \times 10^8$ ms^{-1} in free space. They differ in wavelength (and hence in frequency; $\nu = c/\lambda$) only, which means that the sources that produce them and their detectors are different. Table 9.1 shows the frequency range, wavelength range, the names and the sources of the known electromagnetic radiations. The spectrum of electromagnetic radiations has no upper or lower limits and all the regions overlap.

Table 9.1 The electromagnetic Spectrum

Name	Frequency range (Hz)	Wavelength range (m)	Source
Radiowaves	10^4 to 10^8	0.1 to 600	Oscillating electric circuits
Microwaves	10^9 to 10^{12}	10^{-3} to 0.3	Oscillating currents in special vacuum tubes
Infrared	10^{11} to 5×10^{14}	10^{-6} to 5×10^{-3}	Outer electrons in atoms and molecules
Visible light	4×10^{14} to 7×10^{14}	4×10^{-7} to 8×10^{-7}	Outer electrons in atoms
Ultraviolet	10^{15} to 10^{17}	1.5×10^{-7} to 3.5×10^{-7}	Outer electrons in atoms
x-rays	10^{18} to 10^{20}	10^{-11} to 10^{-8}	Inner electrons in atoms and sudden deceleration of high energy free electrons
Gamma-rays	10^{19} to 10^{24}	10^{-16} to 10^{-13}	Nuclei of atoms and sudden deceleration of high energy particles

Notice that visible light is only a very small part of the total electromagnetic spectrum. We see that electromagnetic waves have a very wide of wavelengths (and hence of frequencies). Although they are identical in nature, their interaction with matter or their physiological action on living bodies depends on their frequency. Infrared rays are thermal radiations which produce heat, x-rays and gamma rays are highly penetrating, to mention only a few of the effects.

9.6 ELECTROMAGNETIC RADIATION AND EARTH'S ATMOSPHERE

The primary source of electromagnetic radiation that we on earth receive is the sun. The solar radiation passes through the different

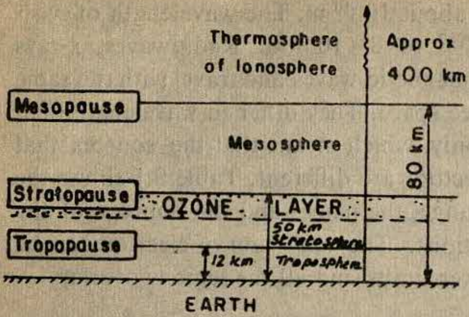


Fig. 9.10 *Various layers of earth's atmosphere*

layers of the earth's atmosphere before it reaches the earth. In fact only a very small fraction of the total radiation manages to reach the earth.

Figure 9.10 shows the names and altitudes of the various layers of the earth's atmosphere. The boundaries between layers are not sharp and the distances given are approximate. The constitution of different layers is different. Water vapour, for example, is concentrated in the lowest layer, above this layer is a region of ozone gas. The upper atmosphere has a layer called the ionosphere which consists of partially ionized gases.

The atmosphere is transparent to visible radiation. The other components (such as infrared and ultraviolet) of the solar radiation are absorbed in different layers. The infrared radiations are called heat or thermal radiations. They produce heat when they are absorbed by a body. Now most infrared radiation from the sun is absorbed by the atmosphere. Whatever little radiation the earth receives heats the earth. Consequently the earth itself emits radiation, and this radiation is mostly in the infrared region (because the earth is much cooler than the sun; hence from Planck's law, it can emit only long-wavelength radiation such as infrared). The infrared radiation emitted from the earth cannot cross the lower atmospheric layers which reflect it back. The clouds also prevent the radiation from escaping and thus act as "blankets" of the earth. The lower atmosphere, therefore, becomes richer in infrared radiation and all objects on earth absorb it getting heated in the process. This phenomenon is called the "Greenhouse effect".

The components of the solar radiation whose wavelength is shorter than the violet (such as ultraviolet and x-rays) are dangerous; they damage the living cells. Fortunately, the ozone layer effectively absorbs these harmful components and we are protected from them.

Radiowave and Microwave Transmission

Radiowaves and microwaves (i.e. waves whose wavelength is longer than about 1 mm) are relevant as far as transmission and communication are concerned. The earliest experiments of Hertz and Marconi on radio communication used waves of wavelengths 10 m or more. The lower atmosphere is transparent to waves of these wavelengths but the upper atmospheric layer called the ionosphere reflects them back to earth. This property is, therefore, made use of in radio communication.

A radio signal emitted from a certain location on earth can be received at another location in two possible ways. The wave can travel directly along the earth from one point to another—this is called the *ground wave*—or it could be directed towards the sky and reflected from the ionosphere towards a desired location on earth—this is called the *sky wave*.

The communication by ground wave is possible with waves of wavelengths longer than about 200 m because the lower wavelength waves are heavily damped by interaction with matter. Below this wavelength, communication through the sky wave is the only method. In common day language the two regions are referred to as the medium wave band and the short wave band.

Communication via the sky wave is not possible if the frequency of the wave is higher than about 40 MHz because the ionosphere bends these waves and does not reflect them. Therefore, television signals (whose frequencies are in the range 100–200 MHz) cannot be transmitted via the sky wave; only ground wave transmission is possible. The receiver antenna is placed directly in the path of the TV signal. The range of a TV signal can be easily computed as shown in the following.

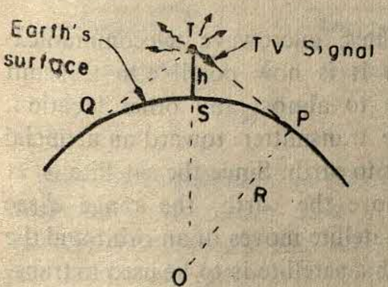


Fig. 9.11 *Range of TV transmission*

Suppose a TV transmitter is stationed on top of a tower at a a height *h* above the ground (Fig. 9.11). Because of the curvature of the earth, the TV signal cannot be received beyond places marked P and Q in the figure. The distance PS or QS up to which signals can be received can be easily calculated in terms of height *h* and the radius *R* of the earth. In the right-angled triangle OTP, we have

$$(OT)^2 = (OP)^2 + (TP)^2$$

where OT $= R + h$, OP $= R$, TP $\simeq$ SP $= d$(say). We then have

$$(R+h)^2 = R^2 + d^2$$

or
$$d^2 = h^2 + 2hR = 2hR\left(1 + \frac{h}{2R}\right)$$

since $h \ll R$, we have

$$d = \sqrt{2hR} \tag{9.38}$$

If the TV transmitter is at a height $h = 50$ m, the value of d works out to 80 km (radius of earth $= 6.4 \times 10^6$ m). The actual range is, in fact, less than 80 km due to heavy attenuation (damping) of the signals unless the transmitter is very powerful, i.e. the signals have a large amplitude (E_0) and therefore a large energy.

Electromagnetic waves of frequencies higher then TV signals are called the microwaves. Their wavelength is of the order of a few millimetres. We will learn in the next chapter that such short wavelength waves are not diffracted (i.e. bent) by obstacles of normal dimensions; they travel in straight lines. Therefore, a beam of microwaves can be sent in a particular direction. A radar is a device that does this; it sends a beam of microwaves towards a distant object and receives the signal reflected back from it. Since the speed of all electromagnetic waves is $c = 3 \times 10^8$ ms^{-1}, from the measurement of the time interval between the sending and receiving of the signal, one can compute the distance of the object.

With the advent of artificial satellites, microwave telecommunications have undergone a revolution. It is now possible to transmit signals from one location on earth to almost any other location. Signals are sent from a microwave transmitter toward an artificial earth satellite which sends them back to earth. Since the satellite is at a great height ($h \sim 36{,}000$ km) above the earth, the range d[see Eq. (9.38)] becomes very large. The satellite moves in an orbit and the earth spins about its own axis. If such a satellite is to be used to transmit signals to a definite region, its orbit must be such that, relative to the earth's surface, it must appear stationary. For a satellite to appear fixed at a place above a certain place on the earth, it must corotate with the earth so that its orbital period round the earth is exactly equal to the rotational period of the earth about its axis of rotation. Such a satellite is called a *geostationary* or *synchronous* satellite. It is easy to calculate what height it should have to be synchronous. It turns out that, for a satellite to appear stationary, it must be placed in orbit round the earth at a height of about 36,000 km from the earth's surface. This is often called the *parking orbit* of the satellite. Relay satellites are placed in parking orbits to transmit TV programmes from one part of the world to another.

SUPPLEMENT

WIRELESS COMMUNICATION

We have seen that radiowaves and microwaves must be used to transmit signals. Therefore speech or music (which are in the audio frequency range of 20 Hz to 20 kHz) must first be converted into radiowave signals. This is what is done at a transmitting station where amplifiers are used to increase the power of the signals. These boosted radio signals are then radiated in space by an antenna. Thus a transmitting station has a microphone (to convert audio signals to electrical signals), an amplifier and an antenna. The receiving station has a receiving antenna, an amplifier (to amplify the weak signals received) and a loudspeaker (to convert electrical signals back to audio signals). The scheme is shown in Fig. 9.12.

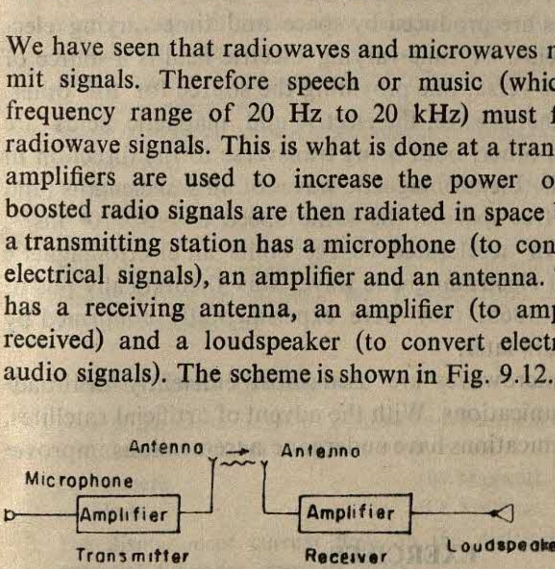


Fig. 9.12

The scheme presented in Fig. 9.12 will not work if the transmitter sends out a harmonic wave of a single wavelength (or a single frequency). The reason is that such a wave travels unchanged; each cycle just like the preceding cycle. If we wish to send a message (speech or music) with the wave, we must bring about a change (or *modulation*) in the wave at the transmitting station in such a way that it can be decoded (or *demodulated*) at the receiving station. Two different modulations are possible. We may modulate either the amplitude or the frequency of the wave. These two ways are respectively called *amplitude modulation* (AM) or *frequency modulation* (FM). The process of modulation at the transmitting station and demodulation at the receiving station is shown in a block diagram (see Fig. 9.13).

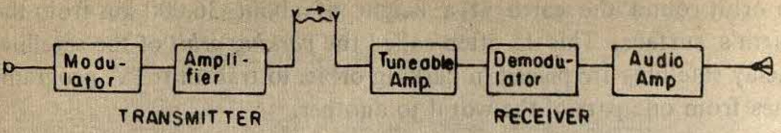


Fig. 9.13

SUMMARY

Electromagnetic waves are produced by space and time-varying electric and magnetic fields. A time-varying electric field is a source of magnetic field and vice versa, one providing the source for the other. In a plane electromagnetic wave, the electric and magnetic fields are at right angles to each other; both being transverse to the direction in which the wave travels. Electromagnetic waves of all wavelengths (and frequencies) travel in free space at the same speed $c = 3 \times 10^8$ ms^{-1}, the speed of light. An accelerated charge emits an electromagnetic wave. The existence of electromagnetic waves was theoretically postulated by Maxwell in 1865, they were experimentally confirmed by Hertz some twenty years later.

Radiowaves and microwaves are used most efficiently for broadcasting and telecommunications. With the advent of artificial satellites, microwave telecommunications have undergone a tremendous improvement in range.

EXERCISES

A. Short-Answer Questions

1. What is the SI unit of the displacement current?
2. What is displacement current due to?
3. Does this current satisfy the continuity requirement.
4. Write down Ampere's law as generalized by Maxwell.
5. What oscillates in an electromagnetic wave?
6. What is a plane wave?
7. Are electromagnetic waves transverse or longitudinal?
8. Is the light from your bulb an electromagnetic wave?

B. Long-Answer Questions

1. Explain clearly how Maxwell was led to the existence of a displacement current.
2. "Conduction and displacement currents individually are discontinuous, but their sum is continuous". Comment.

3. A capacitor with air as dielectric has circular plates, each of radius r and separated by a distance d. It is being charged by an external source; the charging current being I. Obtain the expression for the magnitude of magnetic field at distance R from the axis of the plates if: (a) $R < r$ and (b) $R > r$. Use generalized Ampere's Law. Verify your results using the Biot-Savart law.
4. Write down mathematical formulations of the generalized Ampere's law and Faraday's law of induction for free space and hence show that: (a) a plane polarized electromagnetic wave is transverse, and (b) the speed of the wave is equal to that of light. What conclusion would you draw from this?
5. How are electromagnetic waves produced?
6. Obtain expressions for the average energy densities of the electric and magnetic fields of a plane electromagnetic wave travelling in one direction and show that they are equal.
7. Briefly outline Hertz's experiment. What important conclusion did he draw from his experiment?
8. Name the various known electromagnetic radiations. State the range of their frequencies.
9. What are infrared radiations? How are they useful?
10. What is the "Greenhouse effect"?
11. Write an essay on radiowave and microwave communication.
12. What do you understand by the term "modulation"? What are amplitude and frequency modulations? Why is modulation of a wave necessary at the transmitting station?

C. Multiple-Choice Questions

Choose the correct answers or answers from the given alternatives.

1. Displacement current was first postulated by

(a) Ampere (b) Maxwell
(c) Hertz (d) Marconi

2. The displacement current flows in the dielectric of a capacitor when the potential difference across its plates

(a) is increasing with time (b) is decreasing with time
(c) has assumed a constant value (d) becomes zero.

3. The SI unit of the displacement current is

(a) H (b) C
(c) $F\ m^{-1}$ (d) A

4. Choose the only *wrong* statement from the following. Electromagnetic waves:

(a) are transverse
(b) travel in free space at the speed of light
(c) are produced by accelerating charges
(d) travel with the same speed in all media.

5. Which of the following pairs of space and time varying E and B fields would generate a plane electromagnetic wave travelling in the z-direction?

(a) E_x, B_y (b) E_y, B_x
(c) E_x, B_z (d) E_z, B_x

6. Choose the waves relevant to telecommunication.
   (a) ultraviolet (b) infrared
   (c) microwaves (d) visible light

D. Numerical Problems

1. A air-filled capacitor has circular plates, each of radius 5.0 cm. If the electric field in the air gap changes at a rate of 10^{12} Vm^{-1} s^{-1}, what is the displacement current?
2. What is the magnetic field at each plate of the capacitor of problem 1?
3. A parallel-plate capacitor of area 50 cm^2 and plate separation 3.0 mm is charged initially to 80 μC. Due to a radioactive source nearby, the medium between the plates gets slightly conducting and the plate loses charge initially at the rate of 1.5×10^{-8} Cs^{-1}. What is the magnitude and direction of displacement current? What is the magnetic field between the plates?
4. The voltage between the plates of a parallel-plate capacitor of capacitance 1.0μF is changing at a rate of 5 Vs^{-1}. Find the displacement current in the capacitor.
5. An air-filled capacitor is made of identical circular plates separated by 5.0 mm. Its capacitance is 80 pF. It is being charged by a constant current. If the magnetic field at a distance of 15 cm from the axis of the plates is 2.0×10^{-7} T, find the charging current.
6. The capacitor of Problem 5 is connected to a (220 V, 50 Hz) ac power supply. Find the maximum value of **B** at a point 6.0 cm from the axis between the plates.
7. Electromagnetic waves travel in a medium at a speed of 2.0×10^{8} ms^{-1}. The relative permeability of the medium is 1.0. Find its relative permittivity.
8. In a plane electromagnetic wave of frequency 1.0×10^{12} Hz, the amplitude of the magnetic field is 5.0×10^{-6} T. (a) Find the amplitude of the electric field. (b) What is the total average energy density of the wave?

10

WAVE OPTICS

10.1 INTRODUCTION

The subject of optics has been divided into two branches, namely, *ray optics* and *wave optics.* Ray optics deals with the formation of images in mirrors and lenses, the aberrations of optical images and the working and designing of optical instruments. You have studied this branch of optics in earlier classes. In the next chapter we will learn more about ray optics. Ray optics is based on the laws of rectilinear propagation, reflection and refraction of light and the knowledge of geometry of straight lines. In the next chapter we will see that many aspects of the behaviour of light can be well described and understood in terms of the idea of rays.

During the first decade of the nineteenth century, some more facts about light were discovered. In 1801, Thomas Young demonstrated *interference* between two beams of light. In 1808, Malus discovered that the light reflected from a glass was *polarized.* Grimaldi in 1665 had already discovered *diffraction* of light. These aspects of the behaviour of light (interference, diffraction and polarization) cannot be explained in terms of the idea of rays. In this chapter we will see how these phenomena are explained in wave optics. In chapter 9 we have learnt that light is an electromagnetic wave in which the electric and magnetic fields vary with space and time resulting in the propagation of the wave.

We begin this chapter by describing the connection between rays and wavefront. In wave optics, the wavefront is the counterpart the ray in ray optics. We will then use the idea of wavefronts to explain interference, diffraction and polarization of light. We will end this chapter with a section on *Doppler effect* for light waves.

10.2 RAYS AND WAVEFRONTS: HUYGENS' PRINCIPLE

We can classify waves as one-dimensional, two-dimensional and three-dimensional. Waves on a string are one-dimensional, since they travel only in one direction, i.e. along the string. Surface water waves travel only on the surface of water, and are hence two-dimensional. Sound and light waves which travel in all directions are three-dimensional waves.

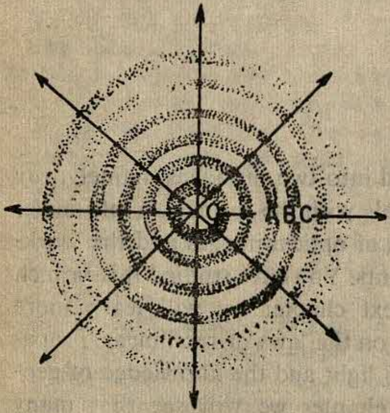


Fig. 10.1 *Wavefronts of water waves*

Let us consider water waves. When a stone is dropped in a still pond, water waves spread out in the form of concentric circles (Fig. 10.1). These circles are called wavefronts. Circle A represents a crest. C is the next crest. Between the circles A and C is circle B which is a trough. All particles in a crest move up together and are at their maximum +ve displacement at the same time. In other words, they have same state of motion, i.e. the phase of oscillation. Similarly, all particles constituting a trough are at their maximum-negative displacement (below the level surface) simultaneously, i.e. they have the same phase. Thus, we define *wavefront* as the locus of points having the same phase of oscillation. In the case of water waves, the loci of points in the same phase are concentric circles, since in a homogeneous medium like water, the disturbance created at O travels the same distance r to reach the crest A. Thus, the phase of particles situated at a distance r from O is the same at a particular instant of time. The ray is the path of an element of the wavefront and hence gives the direction of propagation of the waves.

Let us now consider light waves. The wave travels in all directions. In an isotropic medium light waves created by a point source propagate with the same velocity in all directions. Hence, the wavefronts which envelop points in the same phase of oscillation are spherical in this case and the rays are normal to the wavefronts. The wavefronts of waves travelling in a single direction are plane (Fig. 10.2a). As the spherical wave advances, the curvature of the spheres decreases progressively (Fig. 10.2b). Thus, a small portion of a spherical wave viewed at a long distance from its source is a plane wave, in which case the rays are parallel and unidirectional.

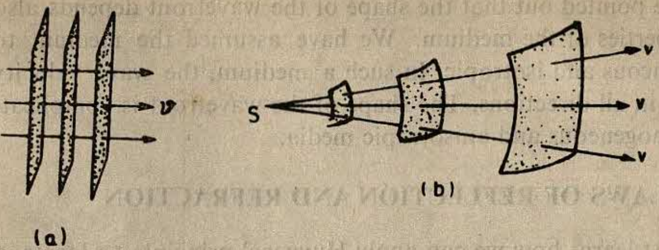


Fig. 10.2 *(a) Plane wavefronts, and (b) Spherical wavefronts; Arrows represent rays*

Huygens' Principle

As a wave spreads out, the wavefronts advance. The radius of circular wavefronts of water waves or of spherical wavefronts of light waves continuously increases. Christian Huygens (1629–1695), a Dutch physicist, give us what we call *Huygens' principle* which tells us how to predict a new wavefront if the position of an earlier wavefront is known. This principle states that each point on a given wavefront is to be considered as the source of secondary wavelets; the new wavefront will then be the envelope of these secondary wavelets. Referring to the circular wavefronts of water waves (Fig. 10.1) let us see how we can use this principle to construct (or predict) a subsequent wavefront out of the given wavefront A. We treat every point on A to be a source (like the original source O) of secondary wavelets. We wish to determine where this wavefront would be after, say, a time interval Δt. The waves will travel a distance $\Delta r = v\Delta t$ in time Δt; where v is the wave velocity. Now, as shown in Fig. 10.3, we draw circles of radius Δr taking each point on A as the centre. We then construct the envelope B of these small wavelets. B is the new wavefront.

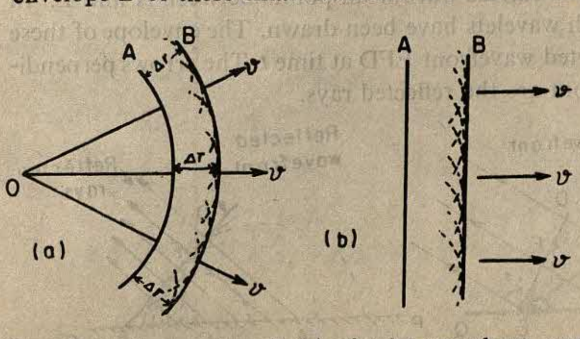


Fig. 10.8 *Propagation of; (a) circular wavefront, and (b) plane wavefront; the arrows indicate the direction of the rays*

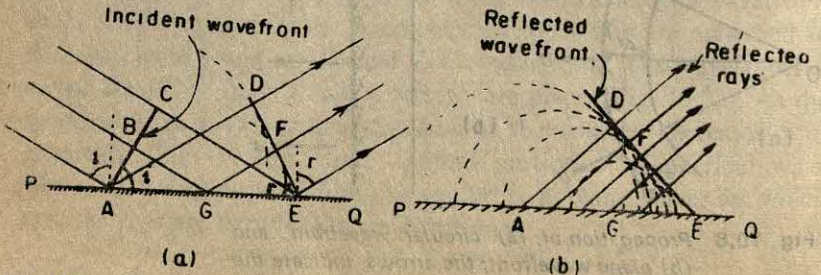
It may be pointed out that the shape of the wavefront depends also on the properties of the medium. We have assumed the medium to be homogeneous and isotropic. In such a medium, the wave velocity is the same in all directions. The shape of the wavefront is complicated in inhomogeneous and anisotropic media.

10.3 THE LAWS OF REFLECTION AND REFRACTION

We have just learnt how we can apply Huygens' principle to locate a given wavefront at a later time. We will now use this principle to deduce the laws of reflection and refraction of a wave.

Laws of Reflection

Let us, for simplicity, consider a plane wave incident on a plane reflecting surface. Let ABC in Fig. 10.4(a) represent a section of a plane wavefront incident on a plane reflecting surface PQ at an angle i. The arrows represent the rays. Let us assume that at time $t = 0$, point A of the wavefront has reached the surface. After some time B will reach G and after some more time C will reach E. Point C will reach E at time t given by $t = v/\mathrm{CE}$, where v is the velocity of the wave. From Huygens', principle, every point of the surface between A and E, in turn, becomes a source of secondary spherical wavelets. Let us now locate the reflected wavefront at time t. At time t, the wavelet that had emerged from A at time $t = 0$, has travelled a distance AD = CE. With A as the centre and AD as the radius, draw a circle (shown only partly). This circle represents the section of the spherical wavelet in the plane of the paper. The wavelet that emerges from G when B reaches G must have travelled a distance GF = BG until time t. We draw another circle with G as the centre and GF as the radius. This process is carried out for all points between A and E. In Fig. 10.4(b) five such wavelets have been drawn. The envelope of these wavelets is the reflected wavefront EFD at time t. The arrows perpendicular to this wavefront are the reflected rays.



(a) (b)

Fig. 10.4 *Reflection of a plane wavefront from a plane surface*

The right triangles ACE and AED are congruent, since CE = AD and AE is the common side.

Hence $$\angle CAE = \angle DEA$$

or $$i = r$$

i.e. the angle of incidence is equal to the angle of reflection. This is the first law of reflection. Further, the incident wave, the reflected wave and the normal at the point of incidence, all lie in the same plane which, in our case, is the plane of the paper. This is the second law of reflection. These laws of reflection hold for all types of waves. You are already familiar with the laws of reflection of light.

Laws of Refraction

Consider a plane wave incident on a plane boundary separating two different media. The wave is partly reflected and partly refracted. We shall consider only the refracted part. Let AB in Fig. 10.5(a) represent a section of a plane wavefront incident at an angle i on a boundary PQ separating two media 1 and 2. Let v_1 and v_2 respectively be the velocity of the incident wave (in medium 1) and the refracted wave (in medium 2). As in the case of reflection, the form and position of the refracted wavefront can be determined using Huygens' construction. Let v_1 be greater than v_2. The first point on the boundary to be hit by the incident wavefront is A and the last is C. From Huygens' principle, every point on the boundary between A and C, in turn, becomes a source of secondary spherical wavelets. Let t be the time taken by the disturbance at B to reach C. Therefore, $BC = v_1t$. During this time the secondary wavelet created at A, t seconds earlier, has travelled a distance $AD = v_2t$, in medium 2. With A as the centre and AD as the radius, draw a circle. In the figure only a part of this

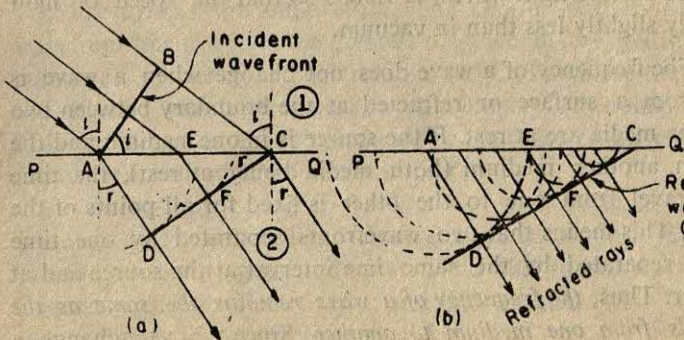


Fig. 10.5 *Refraction of a plane wavefront at a plane boundary between two media*

circle is drawn and shown dotted. This circle represents the secondary spherical wavelet, at time t, that had emerged from A t seconds earlier. Remember the section of a sphere in a plane is a circle. The refracted wavefront CD is then obtained by drawing a tangent from C to this circle. In Fig. 10.5(b), we have drawn the wavelets from six points between A and C and shown that the tangent drawn is indeed the envelope of various secondary wavelets.

Now, in $\triangle$ABC,

$$\sin \angle BAC = \sin i = \frac{BC}{AC}$$

and in $\triangle$ACD,

$$\sin \angle ACD = \sin r = \frac{AD}{AC}$$

$$\therefore \quad \frac{\sin i}{\sin r} = \frac{BC}{AD} = \frac{v_1 t}{v_2 t} = \frac{v_1}{v_2} = {}^1n_2 \qquad (10.1)$$

${}^1n_2$ is called the index of refraction of medium 2 with respect to medium 1. This is Snell's law of refraction. The incident wave, the refracted wave and the normal to the surface, at the point of incidence, all lie in the same plane which is the plane of the paper in our diagram. These laws of refraction hold for all types of waves.

The refractive index of a medium with respect to vacuum is written as n. We then have

$$n = \frac{c}{v} \qquad (10.2)$$

where c is the speed of the wave in vacuum ($c = 3.0 \times 10^8$ ms^{-1} for light waves) and v is the speed of the wave in the medium. The refractive index of air (for light waves) is 1.0003 so that the speed of light in air is only slightly less than in vacuum.

Note: The frequency of a wave does not change when a wave is reflected from a surface or refracted at the boundary between two media, if the media are at rest. If the source is in one medium and the observer in another medium (both media being at rest), the time taken to travel from one to the other is fixed for all points of the wavefronts. This means that two wavefronts separated by one time period are separated by the same time interval at the source and at the observer. Thus, *the frequency of a wave remains the same as the wave travels from one medium to another*. Since $v = \nu\lambda$ a change in speed v implies that the wavelength λ will change as the wave travels from one medium to another.

We have deduced the laws of reflection and refraction for the simplest case of a plane wavefront incident on a plane boundary. These laws hold for all types of wavefronts and for all kinds of boundaries, plane or curved. Figure 10.6(a) shows the reflection of a plane wavefront from a concave mirror and Fig. 10.6(b) shows the refraction of a plane wavefront through a convex lens.

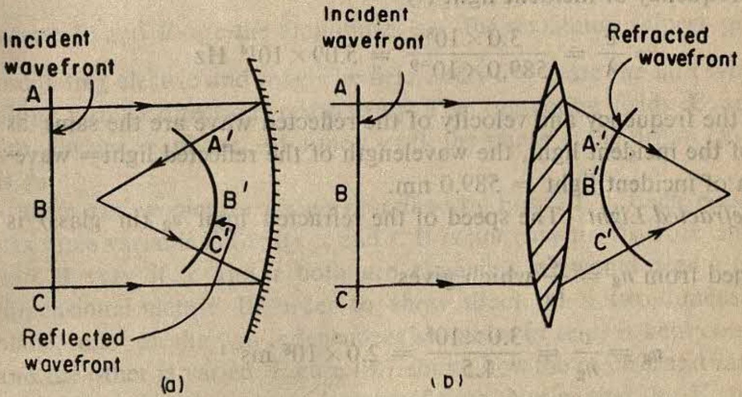


Fig. 10.6 *(a) Reflection of a plane wavefront from a concave mirror, and (b) refraction of a plane wavefront through a convex lens*

Let us first look at Fig. 10.6(b). The central part B of the incident plane wavefront travels through the greatest thickness of the lens and is, therefore, slowed down the most. The outer parts A and C which travel through a smaller thickness of the lens are slowed down the least. This explains the spherical nature of the refracted wavefront and also the converging action of the lens. The concave mirror in Fig. 10.6(a) has the same effect. The central part B of the incident wavefront has to travel the greatest distance before getting reflected, compared to the outer portions A and C. Therefore the point B′ of the reflected wavefront is closer to the mirror than points A′ and C′. This explains the spherical nature of the reflected wavefront and the focussing action of the concave mirror.

EXAMPLE 10.1 Yellow light of wavelength 589.0 nm is incident on a boundary separating air and glass. Find the speed, wavelength and frequency of: (a) reflected light, and (b) refracted light. The refractive index of glass for yellow light = 1.5.

Solution: (a) *Reflected Light* The reflected light is sent back into the medium (air) in which the incident light travels. Now the speed of light in a given medium is fixed $= 3.0 \times 10^8$ ms^{-1} for air (or vacuum). Hence the speed of the reflected light $(c) = 3.0 \times 10^8$ ms^{-1}.

$$\text{Wavelength of incident light } (\lambda) = 589.0 \text{ nm}$$
$$= 589.0 \times 10^{-9} \text{ m}$$

$\therefore$ Frequency of incident light (ν)

$$= \frac{c}{\lambda} = \frac{3.0 \times 10^8}{589.0 \times 10^{-9}} = 5.09 \times 10^{14} \text{ Hz}$$

Since the frequency and velocity of the reflected wave are the same as that of the incident light, the wavelength of the reflected light = wavelength of incident light = 589.0 nm.

(b) *Refracted Light* The speed of the refracted light v_g (in glass) is obtained from $n_g = \frac{c}{v_g}$ which gives

$$v_g = \frac{c}{n_g} = \frac{3.0 \times 10^8}{1.5} = 2.0 \times 10^8 \text{ ms}^{-1}$$

Since the frequency ν of light remains unchanged on refraction, we have

$$\nu = \frac{v_a}{\lambda_a} = \frac{v_g}{\lambda_g}$$

where $v_a = c = 3.0 \times 10^8$ ms^{-1} and $\lambda_a = 589.0 \times 10^{-9}$ m

Therefore

$$\lambda_g = \frac{v_g}{v_a} \cdot \lambda_a$$
$$= \frac{2.0 \times 10^8}{3.0 \times 10^8} \times 589.0 \times 10^{-9}$$
$$= 392.7 \times 10^{-9} \text{ m} = 392.7 \text{ nm}$$

10.4 LIGHT—AN ELECTROMAGNETIC WAVE

In Chapter 9 we have seen that light is an electromagnetic wave in which the electric and magnetic fields vary sinusoidally with space and time. Let us, for simplicity, consider a plane wave. A small portion of any wavefront can be regarded as nearly plane. If a plane wave is travelling in, say, the x-direction, the **E** and **B** fields are confined in the transverse direction, i.e. in the y-z plane. Let us assume that the **E** field is in the y-direction and **B**-field in the z direction. The electric field **E** and magnetic field **B** at a point (x, y, z) at a time t are written as:

$$\mathbf{E}(x, y, z, t) = E_0 \hat{y} \sin \left(\frac{2\pi t}{T} - \frac{2\pi x}{\lambda} + \phi_0 \right)$$

$$\mathbf{B}(x, y, z, t) = B_0 \hat{z} \sin \left(2\pi \frac{t}{T} - \frac{2\pi x}{\lambda} + \phi_0 \right) \qquad (10.3)$$

where E_0 and B_0 are the amplitudes (i.e. the maximum values) of the oscillating electric and magnetic field and $\hat{y}$ and $\hat{z}$ are the unit vectors along y and z directions respectively. The oscillating fields **E** and **B** are always in phase. At time $t = 0$, the phase at the origin $(x = 0)$ is ϕ_0.

How can we picture a wave described by Eqs (10.3). Each equation has three variable **E** (or **B**), x and t. It is not easy to visualize how **E** and **B** vary if x and t both are changing; we would need a three-dimensional picture. In order to show them on a two-dimensional plane, one of the two independent variables (x or t) is kept constant and the other is varied. Figure 10.7 shows how the electric and magnetic fields vary with x at one instant of time. Notice that the **E** and **B** vectors are in the y-z plane which is perpendicular to the x-direction, the direction of propagation. This means that the waves are transverse. The **E** and **B** fields vary as x is changed. If x is changed by λ, their original values are recovered. This is the definition of wavelength λ. Figure 10.7 shows the fields when x is changed by $\lambda/2$. Notice that there is no dependence of **E** and **B** on y and z [see Eqs (10.3)]. This means that the planes parallel to the y-z plane are the wavefronts. Figure 10.7 shows the plane wavefront at five different locations (i.e. different x values). On a wavefront the phase of the wave is constant.

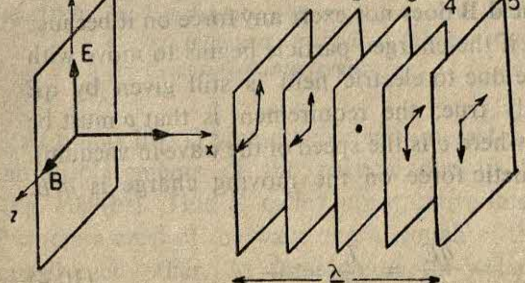


Fig. 10.7 *A plane em wave propagating along the +x-direction*

A similar picture can be drawn keeping x fixed and changing the time t, this picture will give snap-shot of the wave at different times. The planes 1 and 5 now will be separated by a time interval $T/2$.

Charged Particles and Electromagnetic Waves

We know that an accelerating charge radiates electromagnetic waves. Figure 10.8 shows the acceleration vector **a** of a charge at a certain instant of time, the charge is shown accelerating in a direction θ with the x-axis. Let us suppose that an observer is located on the x-axis.

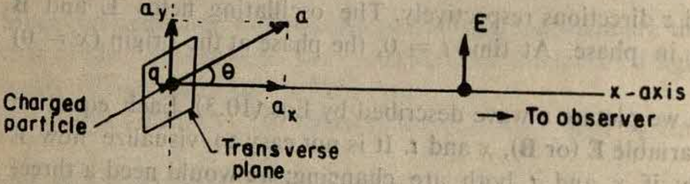


Fig. 10.8 *Radiation from an accelerated charge being studied by an observer inclined at an angle θ with the acceleration vector* **a**

The acceleration vector **a** can be resolved into two rectangular components—one in the direction of the observer (a_x) and the other transverse to this direction (a_y). The component a_x can produce only a longitudinal wave along the x-direction. But electromagnetic waves are transverse, hence the component a_x has nothing to do with electromagnetic waves in the x-direction. Hence the electric vector **E** of the wave is parallel to a_y and proportional to it. We will use this principle later in this chapter when we discuss polarization of light.

Let us, for the present, look at another aspect of electromagnetic waves. Suppose the wave encounters a charged particle at rest. The electric field **E** of the wave will exert a force $q\mathbf{E}$ on the particle if its charge is q; the magnetic field **B** does not exert any force on it because it is at rest. What happens if the charged particle begins to move with a velocity, say, **v**. The force due to electric field is still given by qE (this is only approximately true; the requirement is that v must be very small compared to c; where c is the speed of the wave in vacuum). The magnitude of the magnetic force on the moving charge is qvB. Hence

$$\frac{\text{Electric force}}{\text{Magnetic force}} = \frac{qE}{qvB} = \frac{E}{vB} = \frac{c}{v} \qquad (10.4)$$

where the ratio $E/B = c$ (see Chapter 9).

Thus we find that the electromagnetic wave exerts a much larger electric force (than magnetic force) on a slowly moving charged particle, i.e. if $v \ll c$. In our further discussion, we will therefore consider only the electric field, though the magnetic field is always understood to be present.

10.5 INTERFERENCE

In Class XI you have studied interference of surface water waves which are transverse. In the laboratory this interference can be demonstrated by attaching a stylus to each prong of a tuning fork and placing it over a big ripple tank. The fork is set into vibration and each stylus is brought into contact with water. Two sets of concentric circular waves start simultaneously from the point of contact of each stylus [Fig. 10.9(a)]. The wave trains travel in the form of crests and troughs on the surface of water. The solid circles represents crests and the dotted ones represent troughs.

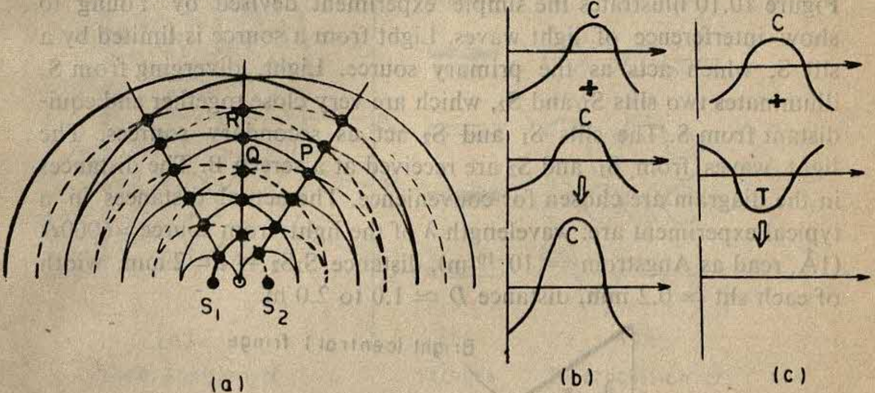


Fig. 10.9 *(a) Interference pattern (solid curves represent crests and broken curves represent troughs), (b) constructive interference, and (c) destructive interference*

At points like Q where two crests interest, a larger crest is formed, the displacements of the particle due to two waves adding up [Fig. 10.9(b)]. This is called constructive interference. At points like P where a crest of the wave and a trough of the other intersect, they cancel each other. At these points, the water surface is level and the interference is said to be destructive [Fig. 10.9(c)]. At points like R, two troughs intersect, the two negative displacements of the same particle add up to give a bigger trough.

The Superposition Principle

The experiment of interference of water waves clearly suggests that the actual displacement of any particle can be found by simply adding the separate displacements that the individual waves will give it if each wave were present alone. This process of vector addition of displacements of a given particle is called *superposition. The superposition principle states that the resultant displacement of a particle is equal to the algebraic* (*vectorial*) *sum of the individual displacements give to it by two or more waves.*

The same arguments can be used to discuss interference of two light waves, except that in place of particle displacements we use the electric fields of the waves. Historically, the phenomenon of interference of light was discovered by the English scientist Thomas Young in 1801.

Young's Double-Slit Experiment

Figure 10.10 illustrates the simple experiment devised by Young to show interference of light waves. Light from a source is limited by a slit S, which acts as the primary source. Light, diverging from S, illuminates two slits S_1 and S_2, which are very close together and equidistant from S. The slits S_1 and S_2 act as secondary sources. The light waves from S_1 and S_2 are received at a screen B. The distances in the diagram are chosen for convenience. The actual distances in a typical experiment are: wavelength λ of the light from source$\simeq$6000Å (1Å, read as Angstrom $= 10^{-10}$ m), distance $S_1S_2 = d \simeq 2$ mm, width of each slit $\simeq$ 0.2 mm, distance $D \simeq$ 1.0 to 2.0 m.

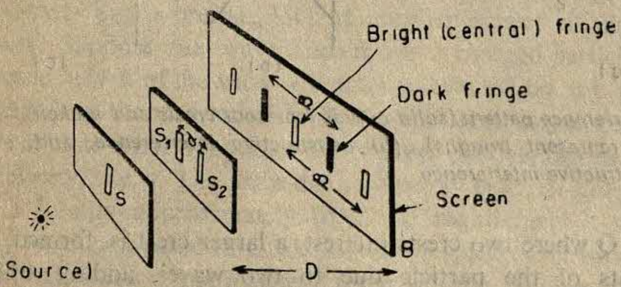
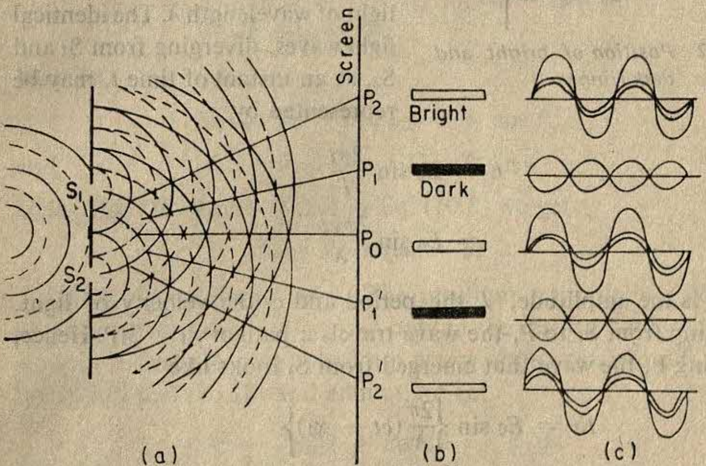


Fig. 10.10 *Young's double slit experiment*

Figure 10.11(a) represents a section of Young's experiment in a plane, i.e. the plane of the paper. According to Huygens' principle, the waves spread out from slit S and as $SS_1 = SS_2$ these waves reach the slits S_1 and S_2 simultaneously, so the waves diverging from S_1 and

S_2 are in the same phase. A train of secondary wavelets diverges to the right from each slit. The solid curves represent crests, i.e. $\mathbf{E}$ + ve maximum, and the dotted curves represent troughs, i.e. $\mathbf{E}$ − ve maximum, in these wave trains. At the lines leading to P_0, P_2 and $\mathbf{P}_2$, the crest of one wave coincides with the crest of the other and the trough of one wave coincides with the trough of the other. Therefore, according to the superposition principle, the electric fields of the waves add up and the resultant amplitude (and hence intensity) is maximum along these lines [Fig. 10.11(c)]. The interference along these directions is constructive. Along the lines leading to P_1 and P_1', the crest of one wave falls over the trough of the other and the effects cancel out, resulting in zero intensity along these directions. This is destructive interference.



(a) Superposition of wavefronts

(b) Fringes

(c) Superposition of fields

Fig. 10.11 *Interference of two light beams*

If the light falls on a screen, alternate bright and dark bands are visible [Fig. 10.11(b)]. These are called interference fringes. The bright fringes are seen to be equally bright and equally spaced. The separation β between the two consecutive bright (or dark) fringes is called the *fringe width.* Fringe width β increases if the distance D, between the slits and the screen, is increased or if the separation d, between the slits, is decreased. The fact that the observed pattern is indeed due to interference of two light waves and not due to any other wave-effect, can be confirmed by covering one of the slits. The well-defined bright and dark fringes will disappear.

Theory of Interference

We shall now obtain the expression for the resultant intensity at any point P of the screen, using the superposition principle (Fig. 10.12).

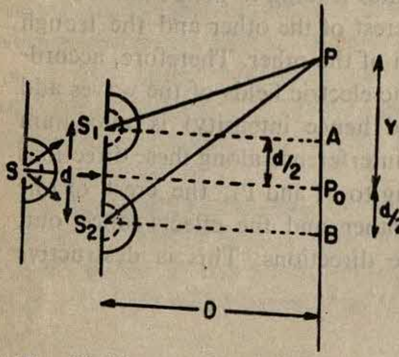


Fig. 10.12 *Position of bright and dark fringes*

Since the two virtual sources S_1 and S_2 are derived from the same primary source and since they are equidistant from S, the secondary wavelets emerging from S_1 and S_2 are in the same phase and have equal amplitude, equal frequency and hence, equal velocity. Let source slit S be illuminated with a monochromatic (of one colour) light of wavelength λ. The identical light waves, diverging from S_1 and S_2, at an instant of time t, may be represented by

$$E = E_0 \sin \frac{2\pi t}{T}$$

$$= E_0 \sin \frac{2\pi ct}{\lambda}$$

where E_0 is the amplitude, T the period and c the velocity of light. In travelling from S_1 to P, the wave travels a path of $x_1 = S_1P$. Hence, on reaching P, the wave that emerged from S_1 looks like

$$E_1 = E_0 \sin \left\{\frac{2\pi}{\lambda}(ct - x_1)\right\}$$

$$= E_0 \sin (\omega t - \phi_1) \qquad (10.5)$$

where $\omega = 2\pi c/\lambda = 2\pi/T$ is the angular frequency and $\phi_1 = 2\pi x_1/\lambda$ is the phase of the wave. Similarly, on reaching P, the wave that emerged from S_2 is given by

$$E_2 = E_0 \sin \left\{\frac{2\pi}{\lambda}(ct - x_2)\right\}$$

$$= E_0 \sin (\omega t - \phi_2) \qquad (10.6)$$

where $x_2 = S_2P$ and $\phi_2 = 2\pi x_2/\lambda$ is the phase of the wave. Notice that the phases of the two waves (which started in phase from S_1 and S_2) are no longer equal. The phase difference ϕ is given by

$$\phi = \phi_2 - \phi_1 = \frac{2\pi}{\lambda}(x_2 - x_1) = \frac{2\pi\Delta}{\lambda} \qquad (10.7)$$

where the path difference $\Delta = x_2 - x_1 = S_2P - S_1P$.

Before we use the superposition principle to find the resultant wave, we will make an assumption. We assume that the point P is very close to P_0, the central point, so that $y \ll D$. We also assume that the two slits are very close together, i.e. $d \ll D$. Under these assumptions, the electric fields E_1 and E_2 at P will be very nearly in the same direction so that the superposition of E_1 and E_2 is simply the ordinary sum of E_1 and E_2. It is no longer necessary to treat them as vectors. Hence the resultant wave is given by

$$\begin{aligned} E &= E_1 + E_2 \\ &= E_0\{\sin(\omega t - \phi_1) + \sin(\omega t - \phi_2)\} \\ &= E_0(\sin\omega t\cos\phi_1 - \cos\omega t\sin\phi_1 + \sin\omega t\cos\phi_2 \\ &\quad - \cos\omega t\sin\phi_2) \\ &= E_0\{(\cos\phi_1 + \cos\phi_2)\sin\omega t - (\sin\phi_1 \\ &\quad + \sin\phi_2)\cos\omega t\} \qquad (10.8) \end{aligned}$$

Now let

$$E_0(\cos\phi_1 + \cos\phi_2) = R\cos\theta \qquad (10.9)$$

and

$$E_0(\sin\phi_1 + \sin\phi_2) = R\sin\theta \qquad (10.10)$$

Using Eqs (10.9) and (10.10) in Eq. (10.8) we get

$$E = R\sin(\omega t - \theta) \qquad (10.11)$$

This is the resultant wave. The resultant amplitude is R and the phase is θ. As we are interested only in the resultant intensity (which is proportional to R^2) we shall not worry about the phase θ. Squaring Eqs (10.9) and (10.10) and adding, we get

$$\begin{aligned} R^2 &= 2E_0^2\{1 + \cos(\phi_2 - \phi_1)\} \\ &= 2E_0^2(1 + \cos\phi) \qquad (10.12) \end{aligned}$$

Constructive Interference

The resultant intensity I ($\propto R^2$) is maximum if R^2 is maximum, i.e. if (see Eq. 10.12)

$$\cos\phi = +1$$

or $$\phi = 0, 2\pi, 4\pi, \ldots$$

or $$\frac{2\pi\Delta}{\lambda} = 0, 2\pi, 4\pi, \ldots$$

or $$\Delta = 0, \lambda, 2\lambda, \ldots$$

The maximum value is $I_{\max} = kR_{\max}^2 = 4kE_0^2 = I_0$

where k is a proportionality constant. The interference is said to be constructive.

Destructive Interference

The resultant intensity is minimum if R^2 is minimum, i.e. if

$$\cos\phi = -1$$

or
$$\phi = \pi, 3\pi, 5\pi, \ldots$$

or
$$\frac{2\pi\Delta}{\lambda} = \pi, 3\pi, 5\pi, \ldots$$

or
$$\Delta = \frac{\lambda}{2}, \frac{3\lambda}{2}, \frac{5\lambda}{2}, \ldots$$

The minimum value is $I_{\min} = 0$.

The interference is said to be destructive. Thus the condition for maxima and minima of intensity are as follows.

$$\Delta = S_2P - S_1P = m\lambda \qquad \text{(MAXIMA)} \qquad (10.13)$$

and

$$\Delta = S_2P - S_1P = \left(m + \frac{1}{2}\right)\lambda \qquad \text{(MINIMA)} \qquad (10.14)$$

where $m = 0, 1, 2, 3, \ldots$.

These maxima and minima constitute the bright and dark fringes.

Position of Bright and Dark Fringes: The Fringe Width

Referring to Fig. 10.12, we see that

$$(S_2P)^2 = (S_2B)^2 + (BP)^2$$

$$= D^2 + \left(y + \frac{d}{2}\right)^2$$

or
$$S_2P = D\left[1 + \frac{\left(y + \frac{d}{2}\right)^2}{D^2}\right]^{1/2}$$

Since we have assumed that $y, d \ll D$, we expand the expression within the square brackets binomially and retain terms of up to the order of $\left(y + \frac{d}{2}\right)^2 \Big/ D^2$. We then get

$$S_2P = D + \frac{1}{2D}\left(y + \frac{d}{2}\right)^2$$

Similarly $\quad S_1P = D + \frac{1}{2D}\left(y - \frac{d}{2}\right)^2$

so that $\quad \Delta = S_2P - S_1P$

$$= \frac{1}{2D}\left[\left(y + \frac{d}{2}\right)^2 - \left(y - \frac{d}{2}\right)^2\right]$$

$$= \frac{yd}{D} \tag{10.15}$$

Bright Fringes For bright fringes, $\Delta = m\lambda$, i.e.

$$y_m \frac{d}{D} = m\lambda$$

or

$$y_m = \frac{m\lambda D}{d} \tag{10.16}$$

which gives the position of the mth bright fringe.

For the central fringe at P_0, $m = 0$ and $y_0 = 0$ and the next bright fringe will be at $y_1 = \lambda D/d$ and so on. For the $(m-1)$th bright fringe, we have

$$y_{m-1} = \frac{(m-1)\lambda D}{d}; \; m = 1, 2, 3, \ldots : \text{etc.}$$

so that the fringe width β is given by

$$\beta = y_m - y_{m-1} = \frac{\lambda D}{d} \tag{10.17}$$

Dark Fringes The dark fringes satisfy the relation

$$y'_m = \left(m - \frac{1}{2}\right)\frac{\lambda D}{d}$$

i.e. the first dark fringe ($m = 1$) is at

$$y'_1 = \frac{1}{2}\frac{\lambda D}{d}$$

the second ($m = 2$) at

$$y'_2 = \frac{3}{2}\frac{\lambda D}{d}$$

and so on. The separation between two consecutive dark fringes is $\beta = \lambda D/d$, the same as that between two consecutive bright fringes. Thus the fringes are equally spaced. Between two bright fringes is a dark one. Fringe width β is directly proportional to D and inversely proportion to d. Further, the intensity of all bright fringes is the same (equal to I_0) and the intensity of dark fringes is zero. These conclusions are in agreement with the experimental observations.

Intensity Distribution Curve

The intensity at any point of the screen is given by [see Eq. (10.12)]

$$I = 2kE_0^2(1 + \cos\phi) = I_0 \cos^2\left(\frac{\phi}{2}\right)$$

where $I_0 = 4kE_0^2$, is the maximum value of I. In terms of y, d, D and λ we have

$$I = I_0 \cos^2\left(\frac{\pi y D}{\lambda d}\right) \tag{10.18}$$

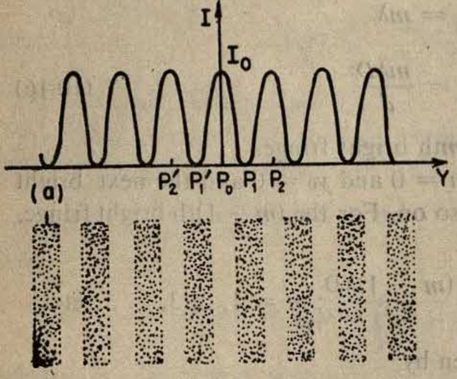


Fig. 10.13 *(a) Intensity distribution, and*
(b) fringe pattern

In Fig. 10.13(a) we have plotted I for various values of y and in Fig. 10.13(b) we show the corresponding fringe system. It may be remarked that there is no destruction of energy at the dark fringes. The total energy is merely redistributed in the fringe system. The energy disappears from the dark fringes and reappears in the bright fringes.

Coherent and Incoherent Light Sources

At first sight, the interference pattern observed on the screen appears paradoxical since we do not observe any pattern on a wall illuminated by two candles. We do not observe any interference pattern if slits S_1 and S_2 in Young's experiment are illuminated by two separate identical monochromatic sources A and B, like two sodium lamps (Fig. 10.14). The reason is not hard to seek. The key word in

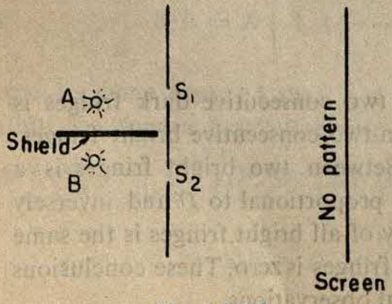


Fig. 10.14

interference is 'phase difference'. We know that the resultant intensity I at any point P, due to the superposition of light waves from sources S_1 aud S_2 is given by

$$I = I_0 \cos^2 \left(\frac{\phi}{2}\right) \tag{10.19}$$

where ϕ is the phase difference between the two interfering waves. This phase difference depends on two factors: (i) the initial phase difference between the sources S_1 and S_2, and (ii) the phase difference resulting from the path difference. At a given point P, the path difference $\Delta = S_2P - S_1P$, is constant. Hence, the phase difference resulting from this path difference does not change with time. If the initial phase difference between the two sources changes with time, the phase difference ϕ in Eq. (10.19) and hence intensity I will change with time. Thus, sustained interference is not possible. This can be understood from the following reasoning. In a light source, namely a glowing filament or a gas discharge tube, light is emitted from individual atoms. The tiniest source may consist of millions of such independent atoms. Each atom is excited by heat or electrical discharge. An electron of the atom shifts to a higher energy orbit and as it falls back to its original orbit, a light is emitted. The duration of this process of emission of light is of the order of 10^{-8} to 10^{-10} s. Millions of such atoms obviously cannot emit light waves in the same phase. The phase of light will change arbitrarily because of collisions with other atoms. Hence, the initial phase difference between two independent sources can remain constant only for about 10^{-8} or 10^{-10} s. Such a rapidly varying phase change (and hence intensity) cannot be seen by us or depicted by any photographic plate, because the time of persistence of vision (0.1 s) is much larger than the so-called time of coherence (10^{-8} or 10^{-10} s). Such independent sources of light which do not have a constant phase difference are called *noncoherent* sources. They cannot produce sustained interference effects.

It is possible to obtain two coherent light sources from one incoherent source. Young's double-slit device is one such example. The two sources (virtual) S_1 and S_2 in Young's experiment (Fig. 10.10) have a constant phase relationship because they are derived from the same source S, which is essentially non-coherent. Whatever phase change takes place in S_1 must also take place in S_2 so that the relative phase difference between S_1 and S_2 remains constant with time. The sources that have a constant phase difference are called *coherent* sources. Only coherent sources can produce sustained or stationary interference. Nowadays we have remarkable devices called *lasers* which produce coherent light.

Fresnel's Biprism

Fresnel devised a simple experiment to obtain coherent sources and produce an interference pattern. The experiment is shown schematically in Fig. 10.15. He used a biprism which is actually a simple prism with very small base angles ($\alpha \simeq \frac{1}{3}°$). The phenomenon of refraction is utilized to produce a pair of coherent sources. Light from a narrow slit S, illuminated with monochromatic light, is allowed to fall on the biprism. The light passes through the biprism. As a result of refraction by the upper and the lower halves of the prism, two virtual images S_1 and S_2 are produced. These virtual sources S_1 and S_2 are coherent as they are derived from the same primary source S. The source S_1 illuminates the portion AB of the screen while the source S_2 illuminates the portion CD. The portion AD receives light from both the sources and interference fringes are seen in this region.

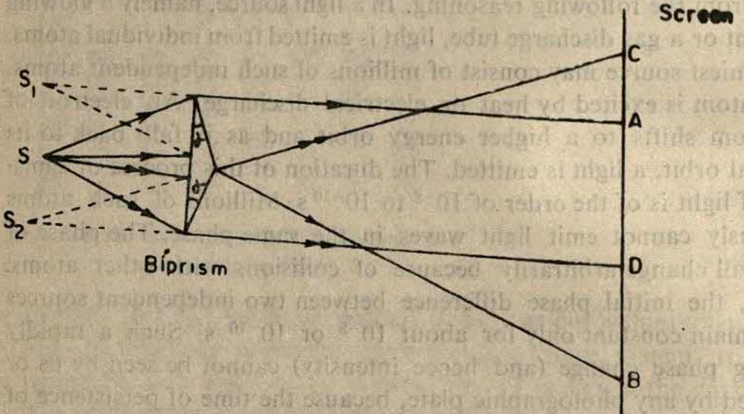


Fig. 10.15 *Fresnel's biprism experiment*

Lloyd's Mirror

In Lloyd' mirror experiment, depicted in Fig. 10.16, the phenomenon of reflection of light is utilized to obtain a virtual light source coherent with its primary source. Light from a slit S_1 is allowed to fall on a plane mirror at grazing incidence (the angle of incidence ≃ 87°). The reflected light appears to diverge from S_2, the virtual image of the source slit S_1, formed by reflection at the mirror. The virtual image S_2 acts as a coherent source with the primary source S_1 itself. S_2 illuminates the region AB of the screen which also receives light directly from S_1. In this region of overlapping of the direct and reflected waves, interference fringes can be seen. It may be mentioned that the conventional silvered mirror is not used because of multiple internal reflections.

In fact, even a plane dielectric surface (like a glass plate) can be employed since, at grazing incidence, glass behaves like a nearly perfect reflector.

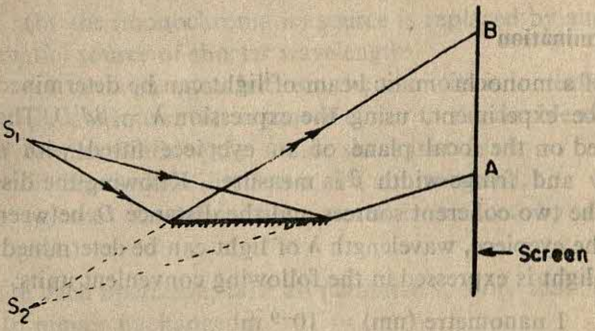


Fig. 10.16 *The Lloyd's mirror*

The shape of fringes and the expression for the fringe width is the same in all the three experiments of interference. An interesting feature of Lloyd's experiment is that the central fringe is dark and not bright. This is to be expected because we know that when a wave is reflected at the boundary of a dense medium, the reflected wave undergoes a phase change of π (which is equivalent to path difference of $\lambda/2$) with respect to the incident wave. Hence, the path difference for any point P on the screen is not just $(S_2P - S_1P)$ but it is equal to $(S_2P - S_1P) + \frac{\lambda}{2}$. Now the condition for destructive interference, Eq. (10.14) becomes

$$S_2P - S_1P + \frac{\lambda}{2} = \left(m + \frac{1}{2}\right)\lambda$$

or $S_2P - S_1P = m\lambda$, where $m = 0, 1, 2, 3, \ldots$ This condition is satisfied for the central fringe for which $S_2P = S_1P$. Thus, the central fringe in Lloyd's experiment should indeed be dark. This provides an experimental confirmation of the fact that the reflected wave undergoes a phase change of π on reflection.

Visibility of Interference Fringes

The separation d between the two virtual sources must be very small in all the three experiments described above otherwise the fringe width β will be very small, i.e. the fringes will be too close together to be clearly visible. It is for this reason that angle α of the biprism is very small or the source slit in Lloyd's mirror experiment is so placed that the light falls on the mirror at almost grazing incidence. Another reason why d as well as y should be small is related to the coherence of the source itself which affects the visibility of the fringes. (Refer to Supplement S1 at the end of this chapter.)

Wavelength Determination

The wavelength of a monochromatic beam of light can be determined by an interference experiment, using the expression $\lambda = \beta d/D$. The fringes are received on the focal plane of an eyepiece fitted with a micrometer screw and fringe-width β is measured. Knowing the distance d between the two coherent sources and the distance D between the sources and the eyepiece, wavelength λ of light can be determined.

Wavelength of light is expressed in the following convenient units:

$$1 \text{ nanometre (nm)} = 10^{-9} \text{ m}$$

$$1 \text{ Angstrom (Å)} = 10^{-10} \text{ m}$$

Thus

$$1 \text{ nm} = 10 \text{ Å}$$

EXAMPLE 10.2 In Young's double slit experiment, the slits are separated by 0.28 mm and the screen is placed 1.4 m away. The distance between the fourth bright fringe and the central bright fringe is measured to be 1.2 cm. Determine the wavelength of light used in the experiment.

Solution: The position of the mth bright fringe with respect to the central fringe is given by

$$y_m = \frac{m\lambda D}{d}$$

For the central bright fringe ($m = 0$), $y_0 = 0$

For the fourth bright fringe ($m = 4$), $y_4 = \frac{4\lambda D}{d}$

Therefore

$$y_4 - y_0 = \frac{4\lambda D}{d}$$

or

$$\lambda = (y_4 - y_0)\frac{d}{4D} \quad \text{(i)}$$

It is given that

$$(y_4 - y_0) = 1.2 \text{ cm} = 1.2 \times 10^{-2} \text{ m}$$

$$D = 1.4 \text{ m}, \ d = 0.28 \text{ mm} = 0.28 \times 10^{-3} \text{ m}$$

Substituting these values in (i) and solving, we get

$$\lambda = 6 \times 10^{-7} \text{ m (or 600 nm or 6000 Å)}$$

EXAMPLE 10.3 What is the effect on the interference fringes in a Young's double-slit experiment due to each of the following operations:

(a) the screen is moved away from the plane of the slits;

(b) the (monochromatic) source is replaced by another (monochromatic) source of shorter wavelength;

(c) the separation between the two slits is increased;

(d) the source slit is moved closer to the double-slit plane;

(e) the width of the source slit is increased;

(f) the widths of the two slits are increased;

(g) the monochromatic source is replaced by a source of white light?

(In each operation, take all parameters, other than the one specified, to remain unchanged.)

Ans. The fringe width is given by

$$\beta = \frac{\lambda D}{d}$$

(a) If the screen is moved away from the plane of the slits, the value of D increases, hence the fringe width β increases.

(b) If wavelength λ is decreased, the fringe width decreased.

(c) If the separation (d) between the two slits is increased, the fringe width decreases.

(d) and (e) Let x be the width of the source slit and X the distance between the source slit and the plane of the two slits. For interference fringes to be distinctly visible, the condition $x/X < \lambda/d$ should be satisfied. If x is too large (i.e. the source slit is too wide) or if X is too small (X is the distance between the source slit and the two slits) the requirement $x/X < \lambda/d$ may be violated and fringes will no longer be distinct. The reason is that the interference patterns due to various parts of the source slit overlap. Consequently, the minima will not be totally dark and fringe pattern becomes indistinct. However, as long as the fringe pattern remains visible, a change in x or X has no effect on the fringe width β.

(f) The single-slit diffraction effects at the two slits will affect the fringe pattern (see Sec. 10.7). The bright fringes are no longer equally bright and equally spaced.

(g) White light consists of colours violet to red (~4000 Å to about 7000 Å). The interference patterns due to different component colours of white light overlap. At the centre of the screen, the path difference is zero for all colours. Hence the central fringe will be white. Since the violet colour has the lowest λ, a violet fringe will be the first fringe on either side of the central white fringe, the farthest fringe is the red fringe. After a few fringes, the visibility of the fringes becomes very poor due to too much overlapping. (See also Supplement S1 at the end of this chapter.)

EXAMPLE 10.4 Monochromatic light of wavelength 6000 Å illuminates two narrow slits 0.3 mm apart producing an interference pattern on a screen 75 cm away. Calculate the separation between: (a) the second bright fringe and the central bright fringe, and (b) the second dark fringe and the central bright fringe.

Solution: (a) The separation between the *m*th bright fringe and the central fringe is given by

$$y_m = \frac{m\lambda D}{d}$$

$$\therefore \quad y_2 = \frac{2\lambda D}{d}$$

$$= 2\times(6000\times 10^{-10})\times\frac{75\times 10^{-2}}{0.3\times 10^{-3}}$$

$$= 3\times 10^{-3}\ \text{m} = 3\ \text{mm}$$

(b) The separation between the *m*th dark fringe and the central fringe is given by

$$y'_m = \left(m - \frac{1}{2}\right)\frac{\lambda D}{d}$$

with $m = 2$, it gives

$$y'_2 = 2.25\ \text{mm}$$

EXAMPLE 10.5 In Young's double slit experiment the slits are 0.5 m apart and interference is observed on a screen placed at a distance of 100 cm from the slits. It is found that the 9th bright fringe is at a distance of 8.835 mm from the second dark fringe from the centre of the fringe pattern. Find the wavelength of light used.

Solution: The distance of the *m*th bright fringe from the central fringe is

$$y_m = \frac{m\lambda D}{d} = m\beta$$

where $\beta = \lambda D/d$ is the fringe width.

$$y_9 = 9\beta \qquad \text{(i)}$$

The distance of the mth dark fringe from the central fringe is

$$y'_m = \left(m - \frac{1}{2}\right)\frac{\lambda D}{d} = \left(m - \frac{1}{2}\right)\beta$$

$$y'_2 = \frac{3}{2}\beta \tag{ii}$$

From (i) and (ii), we get

$$y_9 - y'_2 = 9\beta - \frac{3}{2}\beta = \frac{15}{2}\beta$$

It is given that $y_9 - y'_2 = 8.835$ mm. Hence

$$\beta = \frac{8.835 \times 2}{15} = 1.178 \text{ mm} = 1.178 \times 10^{-3} \text{ m}$$

Now $\lambda = \dfrac{\beta d}{D}$

Substituting for β, d and D, we get

$$\lambda = 5.89 \times 10^{-7} \text{ m} = 589 \text{ nm}$$

EXAMPLE 10.6 In Young's double slit experiment, the intensity of light at a point on the screen where the path difference is λ is K units; λ being the wavelength of light used. Find the intensity at a point where the path difference is $\lambda/4$.

Solution:

$$\text{Phase difference} = \frac{2\pi}{\lambda} \times (\text{path difference})$$

or

$$\phi = \frac{2\pi \Delta}{\lambda}$$

The intensity at a point on the screen is given by

$$I = 2I_0(1 + \cos\phi)$$

where I_0 is the intensity of each interfering beam. At a point where $\Delta = \lambda$, $\phi = 2\pi$ Hence, at this point,

$$I = 2I_0(1 + \cos 2\pi) = 4I_0 = K \text{ units (given)}$$

Thus $I_0 = \dfrac{K}{4}$ units

Therefore $I = \dfrac{K}{2}(1 + \cos\phi)$

When $\Delta = \frac{\lambda}{4}$; $\phi = \frac{\pi}{2}$ Therefore, at this point,

$$I = \frac{K}{2}\left(1 + \cos\frac{\pi}{2}\right) = \frac{K}{2} \text{ units}$$

EXAMPLE 10.7 Two coherent light sources of intensity ratio 25 : 4 are employed in an interference experiment. What is the ratio of the intensities of the maxima and minima in the interference pattern?

Solution: Let I_1 and I_2 be the intensities of the two coherent beams and A_1 and A_2 their respective amplitudes. Now

$$\text{Intensity ratio } \frac{I_1}{I_2} = \frac{25}{4}$$

$$\therefore \quad \text{Amplitude ratio } \frac{A_1}{A_2} = \sqrt{\frac{25}{4}} = \frac{5}{2}$$

i.e. $\quad A_1 = 5$ units and $A_2 = 2$ units

At maxima: $A_{\text{max}} = A_1 + A_2 = 7$ units

At minima: $A_{\text{min}} = A_1 - A_2 = 3$ units

Hence

$$\frac{I_{\text{max}}}{I_{\text{min}}} = \frac{A^2_{\text{max}}}{A^2_{\text{min}}} = \frac{(7)^2}{(3)^2} = \frac{49}{9}$$

10.6 COLOURS OF THIN FILMS

Make a loop of a wire and dip it in a soap solution and take it out. You will get a thin film of soapy water within the frame of the loop. Allow sunlight to fall on the film and look at the reflected light. The film exhibits brilliant colours. On a rainy day small oil films on water show beautiful colours. This is an example of the most common type of interference that occurs in thin films. We shall give an elementary explanation of this phenomenon.

Consider a film enclosed between two surfaces AB and A′B′ (Fig. 10.17). Light from a source S falls on the face AB at point C where it is partly reflected along CD and partly refracted along CE. The refracted ray CE is partly reflected along EF at the lower surface and partly refracted out. This refracted part is not pertinent to our discussion and is, therefore, not shown in the diagram. The ray EF is partly reflected at F (not shown) and partly refracted out along FG. The waves labelled 1 and 2 are coherent since they are derived from the same parent ray SC. When we place our eye so as to receive the

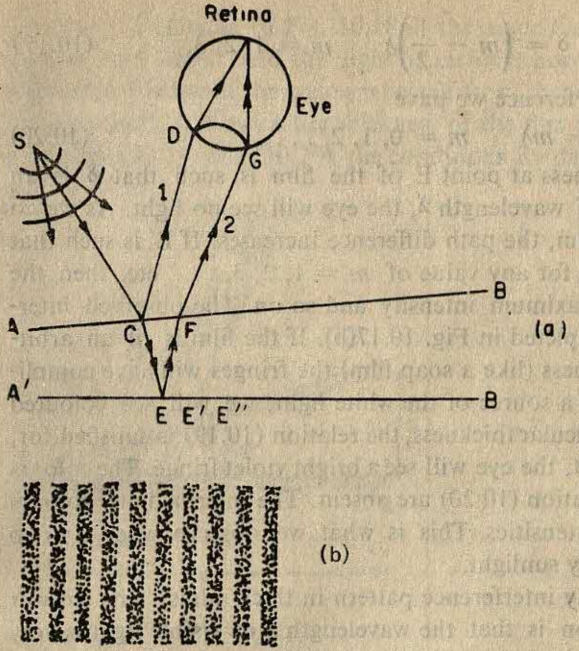


Fig. 10.17 *Interference of light in a thin film: (a) mechanism of interference; and (b) type of interference pattern*

rays 1 and 2, these rays meet each other at the retina of the eye, where they interfere. The observed intensity will depend on the phase difference between waves 1 and 2. The detailed derivation of the phase difference is outside the scope of this book. We shall discuss the phenomenon only qualitatively. Waves 1 and 2 travel different paths to reach the retina. It is obvious from Fig. 10.17(a) that the path difference δ must depend on the thickness of the film at E. Notice that wave 1 is reflected at C and wave 2 at E. If the film is denser than the medium outside it, wave 1 will suffer a phase change of π on reflection while wave 2 will not. A phase change π is equivalent to a path change of $\lambda/2$. Hence, the optical path difference $\varDelta$ between 1 and 2 is given by

$$\varDelta = \delta + \frac{\lambda}{2}$$

where δ depends on the thickness and the angle of incidence and λ is the wavelength of light. Thus, the observed intensity is maximum (constructive interference) if

$$\varDelta = \delta + \frac{\lambda}{2} = m\lambda$$

or
$$\delta = \left(m - \frac{1}{2}\right)\lambda \qquad m = 1, 2, 3, \ldots \qquad (10.19)$$

For destructive interference we have
$$\delta = m\lambda \qquad m = 0, 1, 2, \ldots \qquad (10.20)$$

Now, if the thickness at point E of the film is such that δ is an integral multiple of wavelength λ, the eye will see no light. As we go to point E′ of the film, the path difference increases. If E′ is such that δ for E′ is $(m - \frac{1}{2})\lambda$ for any value of $m = 1, 2, 3, \ldots$ etc. then the eye will register maximum intensity and so on. The observed interference pattern is depicted in Fig. 10.17(b). If the film is of an arbitrarily varying thickness (like a soap film) the fringes will have complicated shapes. If S is a source of the white light, we will see coloured fringes. If, at a particular thickness, the relation (10.19) is satisfied for, say violet component, the eye will see a bright violet fringe. The colours that satisfy the relation (10.20) are absent. The intermediate colours have intermediate intensities. This is what we observe when a soap film is illuminated by sunlight.

Do we observe any interference pattern in thick films? The answer is "no". The reason is that the wavelength λ of visible light is very small. For example, λ of violet light $\simeq 0.00004$ cm, λ of red light is 0.000075 cm, and of green light it is 0.000055 cm. If the thickness t of the film is much larger than λ, the path difference δ will be much larger than λ so that Eqs (10.19) and (10.20) will be satisfied by a large value of the integer m. Since, for large m values, $m \simeq m - \frac{1}{2}$, the maxima and minima of intensity will occur at essentially the same t value. This results in an overlapping of colours, thus, washing out the interference pattern. Hence, to observe the interference pattern, the thickness of the film should be comparable with the wavelength of light.

EXAMPLE 10.8 Figure 10.18 shows two glass plates in contact at one end P and slightly separated at the other end Q, so that a wedge-shaped air film is trapped between them. The film is illuminated normally by monochromatic light and viewed from above. (a) Show that the separation between two successive dark (or bright) fringes is given by $\lambda l/2s$ where s is the thickness of the film at end Q and l is its length. (b) A dark fringe is observed along the line joining the two plates at P. Why? (c) If the wavelength of light is 6000 Å, at what value of the thickness of the film will the first bright fringe be observed starting from P? (d) Is the fringe at a point where the thickness of the film is 0.00015 cm bright or dark? (e) If the space between the plates is filled with water, what changes, if any, do you expect to see in the fringe pattern?

Solution: Referring to Fig. 10.18 if the angle θ of the wedge-shaped film is very small and the light is incident normally on it, the path difference δ between the waves reflected from the top and the bottom faces of the film is twice the thickness of the film at that point. Hence, from Eqs (10.19) and (10.20), the conditions for dark and bright fringes will be

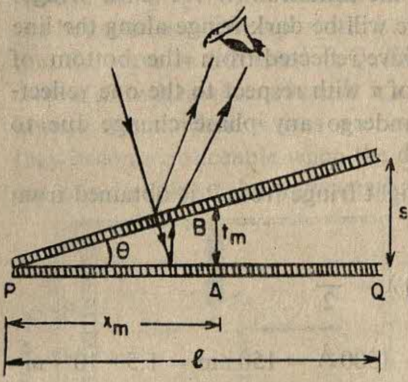


Fig. 10.18

$$2t = m\lambda \text{ (dark fringe)}, \ m = 0, 1, 2, \ldots \qquad \text{(i)}$$

$$2t = \left(m - \frac{1}{2}\right)\lambda \text{ (bright fringe)}, \ m = 1, 2, 3, \ldots \qquad \text{(ii)}$$

(a) Let us suppose that there is mth dark fringe at point A where the film thickness is t_m and let x_m be the distance of that fringe from contact point P. Then in triangle PBA, we have

$$t_m = x_m \tan\theta$$

But $\tan\theta = s/l$. Therefore

$$t_m = \frac{x_m s}{l}$$

Using this in Eq. (i) we have

$$\frac{2x_m s}{l} = m\lambda \qquad m = 0, 1, 2, \ldots$$

or

$$x_m = \frac{m\lambda l}{2s}$$

The $(m - 1)$th dark fringe will be at a point satisfying

$$x_{m-1} = \frac{(m-1)\lambda l}{2s}$$

Therefore, the separation between consecutive dark fringes is

$$\beta = x_m - x_{m-1}$$

$$= \frac{m\lambda l}{2s} - \frac{(m-1)\lambda l}{2s} = \frac{\lambda l}{2s} \qquad \text{(iii)}$$

From Eq. (ii) β for bright fringes is also $\lambda l/2s$.

(b) At point P, $t = 0$. Therefore the condition (i) for dark fringes is satisfied for $m = 0$. Hence there will be dark fringe along the line of contact. The reason is that the wave reflected from the bottom of the film undergoes a phase change of π with respect to the one reflected from the top which does not undergo any phase change due to reflection.

(c) The value of t for the first bright fringe from P is obtained from Eq. (ii) if we set $m = 1$. Thus

$$2t = \left(1 - \frac{1}{2}\right)\lambda = \frac{\lambda}{2}$$

or

$$t = \frac{\lambda}{4} = \frac{6000\ \text{Å}}{4} = 1500\ \text{Å} = 150\ \text{nm} = 1.5 \times 10^{-7}\ \text{m}$$

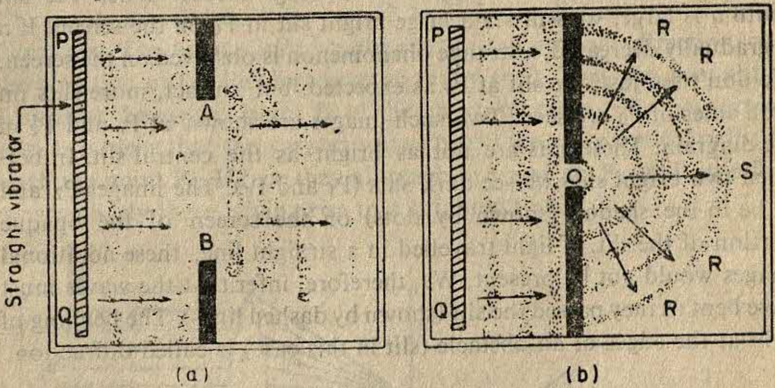
(d) The fringe at $t = 0.00015$ cm will be dark since the relation $2t = m\lambda$ is satisfied for $m = 5$.

(e) Water is denser than air. Hence the wavelength of light will be less in water than in air. From Eq. (iii) it is clear that if λ decreases, β will decrease. Hence the fringes will become closer together when water is introduced between the plates in place of air.

10.7 DIFFRACTION

Another characteristic of wave motion is diffraction which occurs when a travelling wave is obstructed by an obstacle. Let us investigate what happens when an aperture is placed in the path of, for instance, water waves in a ripple tank. Plane waves on the surface of water in a ripple tank can be generated by giving a sideways periodic motion to a straight vibrator PQ (Fig. 10.19). An obstacle, say, an aperture or a slit AB, is placed in the path of the plane waves that are travelling to the right. If the opening AB is much larger than the wavelength of the waves incident on it, the waves pass through it as plane (or nearly plane) waves. The boundaries or the edges of the emergent plane wavefronts correspond to the edges A and B of the slit. The arrows indicate the rays. The wavefronts travelling in straight lines continue to travel in straight lines after emergence from the wide opening AB. We say that the law of rectilinear propagation is obeyed in this case [Fig. 10.19(a)].

A seemingly peculiar phenomenon occurs when the width of the opening AB is decreased so that it is of the order of or less than the wavelength of the waves incident on it [Fig. 10.19(b)]. Parallel plane waves, approaching from the left, pass through the small opening O and spread out into approximately concentric arcs. These circular waves no longer level in straight lines. Their effect is, therefore, experienced not only in the straight direction marked S but also around the boundaries of the aperture, i.e. in the regions marked R. We say that the wave is diffracted at O. *The phenomenon of bending of waves around the edges of an abstacle is called diffraction.* Diffraction effects always occur when waves are incident on openings or obstacles but they become noticeable when the dimensions of openings or obstacles are comparable to the wavelength.



(a) (b)

Fig. 10.19 *Diffraction of water waves: (a) wide aperture, and (b) narrow aperture*

The phenomenon of diffraction is common to all waves. We are all familiar with the diffraction of sound waves. While playing the game of 'hide and seek' we make use of diffraction of sound. You may be hiding in a corner of a building but you could hear the words, 'I spy' uttered by the 'spy' when there is no possibility of his sound reaching you by reflection and refraction. Diffraction of sound is easily observable because of the comparatively long wavelength of sound. For example, the wavelength of an audible sound of frequency 700Hz in air at 25°C is approximately 50 cm. Hence, sound waves are noticeably diffracted at fairly big objects. Diffraction effects of light waves are not so obvious because of their extremely short wavelength. The wavelength of visible light is roughly 6×10^{-5} cm. Hence, diffraction of light can be observed only by allowing light to fall on extremely tiny objects.

You have observed diffraction of light without knowing it. When you look at a distant street lamp through a handkerchief, you see, not one, but a row of points of light arranged symmetrically around the brightest central point. Light is bent as it passes through the various tiny spaces between the strands of the fabric. If you hold up your index finger and middle finger, so that they are barely touching, and look through the small opening at a strong light with your eye focussed at the light, you will be able to see a number of dark lines in the gap between your fingers. These dark lines vanish if the gap is made a little wider. These dark lines are a diffraction pattern due to a single slit. We shall give an elementary explanation of such a diffraction.

Single Slit Diffraction

Let us analyse the diffraction pattern observed on the screen when a plane wave is incident normally on a slit (Fig. 10.20). When the slit width *a* is large, we will see a large bright slit at P_0 on the screen. If *a* is gradually decreased, a strange phenomenon is observed on the screen. We don't see just one slit at P_0 as expected, but, in fact, more slits on both sides of P_0 are seen. Two such images are shown at P_2 and P_2' in the diagram. These slits are not as bright as the central slit. In between two bright slits we see dark slits (P_1 and P_1'). The images P_2 and P_2' lie in the shadow (shown by dots) on the screen of the opaque portion of the slit. If light travelled in a straight line, these additional images would not be present. We, therefore, infer that the waves must have bent as they passed the slit (shown by dashed lines). The bending of light at the edges of an obstacle (slit in this case) is called diffraction.

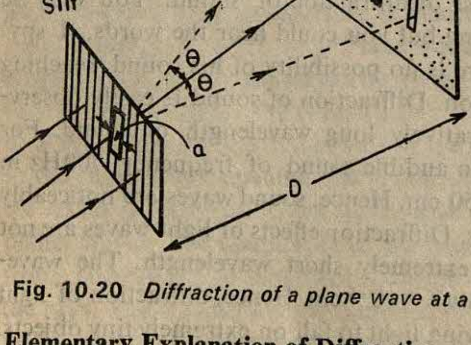


Fig. 10.20 *Diffraction of a plane wave at a slit*

Elementary Explanation of Diffraction

The observed diffraction pattern can be explained by Huygens' principle of secondary wavelets. For simplicity, we shall assume that the distance D between the screen and the slit is very large so that the rays reaching the screen are parallel. The parallel rays can be received on the screen by placing a convex lens (not shown in the figure) between the slit and the screen. Let us imagine that the slit consists of a large number equally spaced points. According to Huygens' principle, each point along the slit becomes a source of secondary spherical wavelets which are all in phase. We shall now examine the directions θ along which constructive and destructive interference can take place.

Straight Direction ($\theta = 0$)

The slit AB, shown in Fig. 10.21, is enlarged for convenience. For the sake of argument the slit of width a is divided into 12 equal parts. The points labelled 0, 1, 2, . . . , 12 send out spherical waves in all directions. The arrows indicate the rays in the straight direction, i.e. the direction of the incident rays, so that $\theta = 0$. They are brought to a focus at the central point P_0 of the screen by means of a converging lens. It is clear that the various waves travel the same optical paths to reach the point P_0. Hence they all reach in phase at P_0 and interfere constructively there, giving rise to a bright slit at P_0.

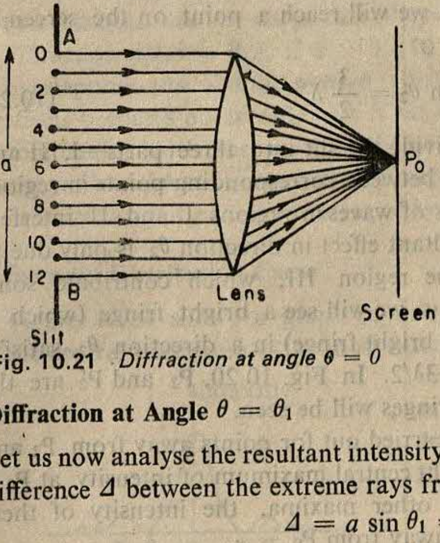


Fig. 10.21 *Diffraction at angle $\theta = 0$*

Diffraction at Angle $\theta = \theta_1$

Let us now analyse the resultant intensity in a direction so that the path difference Δ between the extreme rays from the slit is equal to λ, i.e.

$$\Delta = a \sin \theta_1 = \lambda \tag{10.21}$$

where λ is the wavelength of the incident wave (Fig. 10.22). The path difference between the waves from 0 and 12 is BC, given by $\Delta = \text{BC} = a \sin \theta$. Let us divide the slit into two parts I and II. It is clear from the figure that the waves from 0 and 6 differ in path by $\lambda/2$ and in phase by π. These two waves, therefore, interfere destructively.

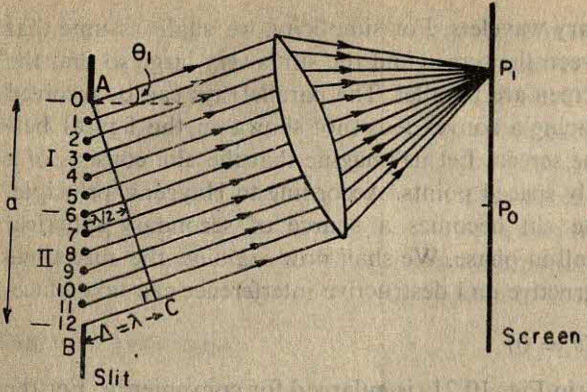


Fig. 10.22 *Diffraction at an angle θ_1, given by $a \sin \theta_1 = \lambda$*

Similarly, the pairs (1, 7), (2, 8), (3, 9) (4, 10), (5, 11), (6, 12) from regions I and II all interfere destructively. Hence, the resultant intensity along the direction θ_1, given by $a \sin \theta_1 = \lambda$ is zero. Thus, we get a dark slit at P_1 (Fig. 10.20). Similarly, we will get a dark fringe at a point P_1 to the right of P_0 since for P_1' also $a \sin \theta_1 = \lambda$.

Diffraction at Angle $\theta = \theta_2$

Moving further away from P_1 we will reach a point on the screen at which the relation

$$\Delta = a \sin \theta_2 = \frac{3}{2} \lambda \qquad (10.22)$$

is satisfied (Fig. 10.23). We divide the slit into three parts—I, II and III. Since the path difference, between corresponding points in regions I and II is again $\lambda/2$, the pairs of waves in regions I and II interfere destructively. Hence, the resultant effect in direction θ_2 is only due to the waves emerging from the region III, which contribute some intensity in this direction. Thus, we will see a bright fringe (which is much fainter than the central bright fringe) in a direction θ_2, satisfying the relation $a \sin \theta_2 = 3\lambda/2$. In Fig. 10.20, P_2 and P_2' are the direction along which these fringes will be seen.

The above analysis can be carried out for points away from P_2 and P_2'. We conclude that the bright central maximum of intensity at P_0 is bordered on both sides by other maxima, the intensity of these maxima decreasing as we go away from P_0.

Let us compute the angular separation between the central maximum ($\theta = 0$) and the first maximum on either side ($\theta = \theta_2$).

Equation (10.22) gives

$$\sin \theta_2 = \frac{3}{2} \frac{\lambda}{a}$$

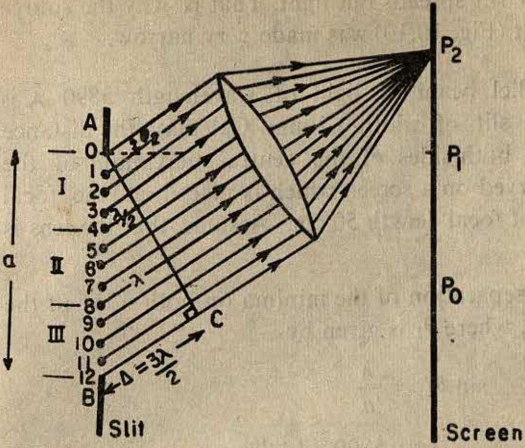


Fig. 10.23 *Diffraction at angle θ_2 given by $a \sin\theta_2 = \frac{3}{2}\lambda$*

For a slit of width $d = 1$ mm illuminated by a typical visible light of wavelength $\lambda = 6000$ Å $= 6 \times 10^{-5}$ cm (yellow light) we have, $\sin\theta_2 = 0.0009$ or $\theta_2 = 1/20$ degree. Thus for a slit 1 mm wide, the two maxima are so close that they nearly overlap, washing out the diffraction pattern. But, if $d = 3 \times 10^{-3}$ cm, $\theta_2 = 17.5°$. There is no overlapping now and the pattern is distinctly visible. Hence, diffraction effects becomes observable only if the dimensions of the obstacle are comparable with the wavelength of light. It is for this reason that x-rays ($\lambda \sim 1$Å) are not easily diffracted while radio waves ($\lambda \sim 10$ m or more) are diffracted by larger objects.

Intensity Distribution Curve

Figure 10.24 shows a graph of intensity versus angle θ. Notice that the central maximum is the most intense. The intensity falls to zero on either side of the central maximum at an angle $\theta = \theta_1$; so that $2\theta_1$ is the angular width of central maximum.

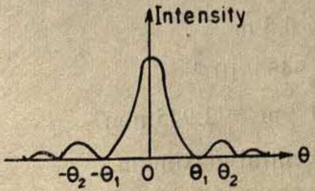


Fig. 10.24 *Intensity distribution in single slit diffraction*

From Eq. (10.21), we find that

$$\sin\theta_1 = \frac{\lambda}{a}.$$

If slit width a is comparable with wavelength λ if $\lambda \simeq a$, $\sin\theta_1 \simeq 1$ so that $\theta_1 \simeq 90°$, i.e. the incident light bends through about

90°. Thus a very narrow slit spreads out light. That is why the source S in Young's experiment (Fig. 10.10) was made very narrow.

EXAMPLE 10.9 A parallel beam of light of wavelength 5890 Å is incident normally on a slit of width 0.2 mm. Calculate the distance between the minima on both sides of the central maximum of the diffraction pattern observed on a screen which is placed at the focal plane of a convex lens of focal length 50 cm. Assume that the lens is placed close to the slit.

Solution: The angular separation of the minima on both sides of the central maximum is $2\theta_1$, where θ_1 is given by

$$\sin\theta_1 = \frac{\lambda}{a}$$

Here $\lambda = 5890\times10^{-10}$ m

and $a = 0.2\times10^{-3}$ m

$$\therefore \quad \sin\theta_1 = \frac{5890\times10^{-10}}{0.2\times10^{-3}}$$

$$= 2.945\times10^{-3}$$

Since θ_1 is small, $\sin\theta_1 \simeq \theta_1$

$$\therefore \quad \theta_1 = 2.945\times10^{-3} \text{ rad}$$

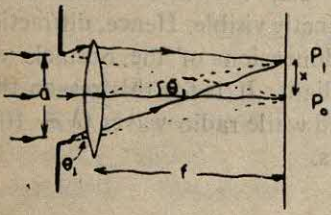


Fig. 10.25

From Fig. (10.25) it is obvious that if the lens is placed close to the slit, $x = f\tan\theta_1 \simeq f\theta_1$, where x is the position of the first minimum (P_1) from the central maximum. Therefore, the distance between the two minima on both sides of the central maximum is

$$2x = 2f\theta_1$$

Now,

$$f = 50 \text{ cm} = 0.5 \text{ m}$$

$$2x = 2\times0.5\times2.945\times10^{-3}$$

$$= 2.945\times10^{-3} \text{ m} = 2.945 \text{ mm}$$

The separation $2f\theta_1$ is the width of the central maximum.

EXAMPLE 10.10 A parallel beam of monochromatic light of wavelength 6000 Å is incident normally on a narrow slit of width 0.3 mm. The diffraction pattern is observed on a screen which is placed at the focal plane of a convex lens placed between the slit and the screen. Calculate the angular separation between (a) the first minimum and the central maximum and (b) the first maxima on both sides of the central maximum.

Solution: (a) The angular separation θ_1 between the central maximum and the first minimum is given by

$$a \sin \theta_1 = \lambda$$

Here, $$a = 0.3 \text{ mm} = 0.3 \times 10^{-3} \text{ m}$$

and $$\lambda = 6000 \times 10^{-10} \text{ m}$$

$$\therefore \qquad \sin \theta_1 = \frac{6000 \times 10^{-10}}{0.3 \times 10^{-3}} = 2 \times 10^{-3}$$

Since θ_1 is very small, $\sin \theta_1 \simeq \theta_1$

$$\therefore \qquad \theta_1 = 2 \times 10^{-3} \text{ rad}$$

(b) The angular separation θ_2 between the central maximum and the first maximum is given by

$$\sin \theta_2 = \frac{3\lambda}{2a} = 3 \times 10^{-3}$$

$$\therefore \qquad \theta_2 = 3 \times 10^{-3} \text{ rad}$$

Since the central maximum is bordered by two maxima symmetrically placed on both sides, the angular separation between the two maxima is $2\theta_2 = 6 \times 10^{-3}$ radian.

EXAMPLE 10.11 In Young's double slit experiment (Fig. 10.12) the distance d between the slits S_1 and S_2 is 2.0 mm. What should the width of each slit be so as to obtain 10 maxima of the two slit interference pattern within the central maximum of the single slit diffraction pattern?

Solution: We have seen that the linear separation on the screen (Fig. 10.12) between m bright fringes in the two-slit interference pattern is given by [Eq. (10.16)]

$$y_m = \frac{m\lambda D}{d}$$

Since $y \ll D$, the angular separation between m bright fringes will be

$$\theta_m = \frac{y_m}{D} = \frac{m\lambda}{d}$$

For 10 bright fringes we have

$$\theta_{10} = \frac{10\lambda}{d} \qquad \text{(i)}$$

Now the angular width of the principal maximum in the diffraction pattern due to a slit of width a is

$$2\theta_1 = \frac{2\lambda}{a} \tag{ii}$$

Equating (i) and (ii), we get

$$\frac{10\lambda}{d} = \frac{2\lambda}{a}$$

or

$$a = \frac{d}{5} = \frac{1.0 \text{ mm}}{5} = 0.2 \text{ mm}$$

We notice that if a is comparable with d, the interference pattern cannot be seen. Hence the two slits in Young's experiment must be very narrow.

The Diffraction Grating

An arrangement of a large number of very narrow equidistant and parallel slits is called a plane diffraction grating. Let N be the number of slits on the grating and d the separation between two successive slits.

Consider a parallel beam of light incident normally on a plane grating. Each slit will diffract light in all directions. If we are looking along a certain direction with respect to the incident direction, the path difference between the waves diffracted by two successive slits is given by (see Fig. 10.26)

$$\Delta = d \sin\theta$$

if $\Delta = n\lambda$, the condition of constructive interference will be satisfied

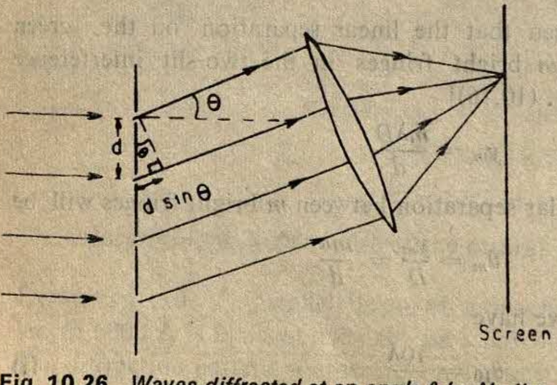


Fig. 10.26 *Waves diffracted at an angle θ by N slits*

between waves diverging from any pairs of slits. Hence waves from all slits will be reinforced in this direction. The condition for this reads

$$d \sin \theta = n\lambda \tag{10.23}$$

where n has values 0, 1, 2, 3, . . .
For $n = 0$ (i.e. $\theta = 0$) we have the central principal maximum. Values of $n = 1, 2, \ldots$ correspond to first order, second order, . . . maxima on both sides of the principal maximum. Their directions are given by:

First order— $$d \sin \theta_1 = \lambda \quad \text{or} \quad \sin \theta_1 = \frac{\lambda}{d} \tag{10.24}$$

Second order— $$d \sin \theta_2 = 2\lambda \quad \text{or} \quad \sin \theta_2 = \frac{2\lambda}{d} \tag{10.25}$$

Figure 10.27 shows the intensities of the different maxima up to the second order. The central principal maximum is the most intense. There are a large number of minima and very weak secondary maxima between any two principal maxima; their number depends on N. The secondary maxima are so weak that they cannot be seen.

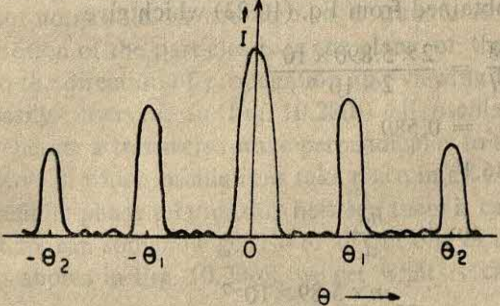


Fig. 10.27 *Intensity distribution in the diffraction pattern due to a grating using monochromatic light*

It is interesting to see what happens if a beam of white light falls on a grating. White light is composed of various colours from violet ($\lambda_v \sim 4000$ Å) to red ($\lambda_R \sim 7000$ Å). It follows from Eq. (10.23) that in a given order n, the value of θ will be different for different colours; θ will be the smallest for violet and largest for red with other colours having intermediate values of θ. This means that the grating will send different colours in different directions. If we use a convex lens and focus them on a screen, as shown in Fig. 10.26, we would obtain the spectrum of white light in very much the same way as in a prism. By measuring the value of θ for each colour, for a given order (say, $n = 1$) and knowing the value of d of the grating, we can determine the wavelength of each colour of white light by using Eq. (10.23). Equation (10.23) is called the *grating equation* and the separation d between two successive slits is called the *grating element*. A grating is made by drawing (with a diamond point) a very large number of straight and equidistant scratches or lines (called *rulings*) over a small width of a thin glass plate. The transparent regions between rulings serve a slits.

EXAMPLE 10.12 Sodium light of wavelength 5890 Å is incident normally on a diffraction grating of 5000 rulings per cm. (a) Find the angle of diffraction for the second order. (b) What maximum order can we get for this grating for sodium light?

Solution: (a) Grating element $(d) = \frac{1 \text{ cm}}{5000} = 2 \times 10^{-4}$ cm

$$= 2 \times 10^{-6} \text{ m}$$

Wavelength $(\lambda) = 5890$ Å $= 5.890 \times 10^{-7}$ m
Order $(n) = 2$

The angle of diffraction is obtained from Eq. (10.23) which gives

$$\sin\theta = \frac{n\lambda}{d} = \frac{2 \times 5.890 \times 10^{-7}}{2 \times 10^{-6}}$$

$$= 0.589$$

giving $\theta = 36.1^\circ$

(b) We have seen that $\sin\theta = \frac{n\lambda}{d}$

$$= \frac{n \times 5.89 \times 10^{-7}}{2 \times 10^{-6}}$$

$$= \frac{n}{3.4}$$

Thus we find that $\sin\theta$ exceeds unity (which is not possible) for $n = 4$. Hence $n = 3$ (i.e. third order) is the maximum order possible for the given grating using sodium light. Orders higher than 3 are possible for lights of shorter wavelengths or with gratings of bigger grating elements.

Validity of Ray Optics

We have seen that the bending of a beam of light due to diffraction at a slit becomes large if the slit width a is comparable with wavelength λ. Since the wavelength of visible light is small (between 4×10^{-7} m to 7×10^{-7} m) most obstacles in many common situations are not small enough (in dimensions) to cause any serious diffraction effects. This means that, under ordinary circumstances, light travels in straight lines and ray optics is valid. Ray optics fails in special situations where the sizes of the obstacles are comparable with wavelength.

10.8 POLARIZATION OF LIGHT

The phenomena of reflection, refraction, interference and diffraction are common to both transverse and longitudinal waves, mechanical as well as electromagnetic. The distinguishing feature is that only transverse waves can be *polarized.* Let us first discuss polarization of a mechanical wave. Let us look at a longitudinal wave and a transverse wave along the direction of propagation.

Imaging the wave to originate from below the page and to travel towards you at right angles to the plane of the page. In the case of a longitudinal wave [Fig. 10.28(a)] the motion of the particles is perpendicular to the page and viewing from above the page this motion is not noticeable. In the case of a transverse wave [Fig. 10.28(b)] the motion of the particles is in the plane of the paper, being transverse to the direction of propagation and viewing from above this motion is easily observed. In Fig. 10.28(b) all oscillations along 11′, 22′, . . . generate a transverse wave perpendicular to the paper. This transverse wave in which oscillations take place in all possible directions with no definite phase relationship between them is called an *unpolarized wave.* If we can somehow get rid of all but one of these oscillations, say 11′, as shown in Fig. 10.28(c), we get what is called a *linearly polarized wave.*

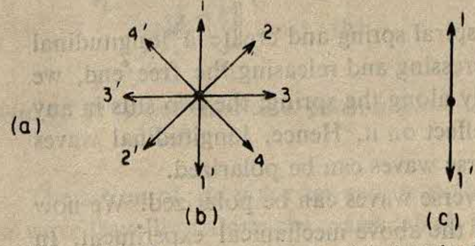


Fig. 10.28 *End-on view of: (a) longitudinal wave, (b) unpolarized transverse wave, and (c) linearly polarized transverse wave*

The following simple experiment helps in understanding polarization, it demonstrates the polarization of transverse waves on a string. Take a long string AB passing through two slits S_1 and S_2, each slit being only slightly wider than the thickness of the string [Fig. 10.29(a)]. The end B is fixed and transverse waves are generated at the free end A by shaking it vertically, horizontally or in any arbitrary manner, so that unpolarized waves, with vibrations in all directions perpendicular to the direction AB, begin to travel along AB. In the portion AS_1, this wave has particle vibrations in all possible planes. In the figure we have shown two such vibration planes. The slit S_1 will permit only those waves to pass that have particle vibrations parallel to the slit S_1. The wave emerging from the slit S_1 is hence polarized. The slit S_1 may be called the polarizing slit or polarizer. To discover that the wave emerging from slit S_1 is polarized we place another slit S_2 called the analyser, so that it is exactly parallel to S_1. The wave will be seen to emerge from it unchanged. If S_2 is now rotated to a position perpendicular to S_1, no vibration at all will emerge from S_2 [Fig. 10.29(b)]. If the analyser slit is horizontal when the wave is cut off, the incoming wave must have been polarized in the vertical plane.

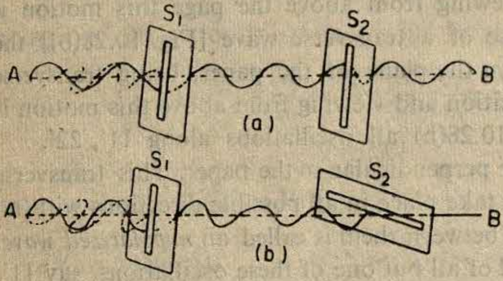


Fig. 10.29 *Mechanical experiment to demonstrate polarization of a wave on a string*

If we replace the string by a spiral spring and create a longitudinal wave in it by periodically compressing and releasing the free end, we will notice that it will pass freely along the spring; the two slits in any relative orientation having no effect on it. Hence, longitudinal waves are not polarized. Only transverse waves can be polarized.

We have seen that only transverse waves can be polarized. We now describe the optical analogue of the above mechanical experiment. In an unpolarized light, the electric field vector has all the possible orientations in a plane perpendicular to the direction of propagation [Fig. 10.28(b)]. If we can get rid of all but one of these orientations, we would get what we called a *linearly polarized* light wave. A specially cut crystal of tourmaline or calcite has just this property. Such a crystal is called a *polaroid* as it serves as the optical analogue of the slit S_1 of Fig. 10.29.

Hold a polaroid P_1 normally to the path of a beam of light from a source S as shown in Fig. 10.30. You will see light transmitted by P_1. Rotate P_1 about the incident beam. No variation in the intensity of emergent beam is observed. Now take another polaroid P_2 and view the light coming out of P_1 through this. Rotate P_2. At a particular orientation of P_2 with respect to that of P_1, the intensity of light emerging from P_2 will be seen to be maximum [Fig. 10.30(a)]. As P_2 is rotated from this position, the intensity of light is seen to decrease till at a certain orientation (90° away from the first orientation) no light is transmitted by P_2 [Fig. 10.30(b)]. A comparison with the mechanical experiment of Fig. 10.29 suggests that the light emerging from P_1 must be polarized. The polaroid P_1 allows only that vibration of electric vector which is parallel to the arrow (↑) of P_1. The light emerging from P_1 is thus polarized. This light has the same orientation of the electric vector as that of P_2 and is, therefore, transmitted by P_2. When P_2 is rotated through 90°, its orientation is perpendicular to that of the light incident on it. Hence, it does not transmit the light. The polaroids are then said to be '*crossed*'. This experiment clearly demonstrates polarization of light waves and confirms that light waves must be transverse.

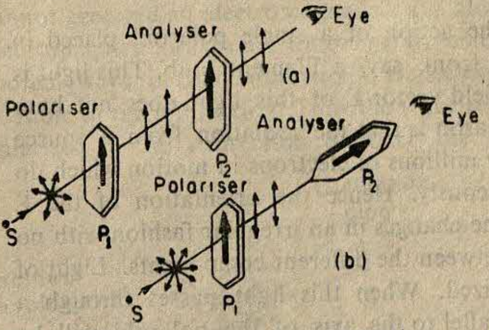


Fig. 10.30 *Demonstration of polarization of light*

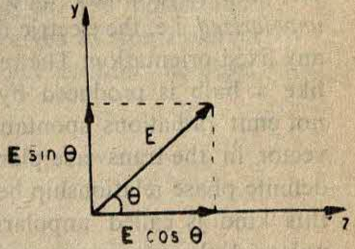


Fig. 10.31 *Action of a polaroid*

The action of the polarizer can be understood as follows. Consider a wave travelling along the x-axis which is perpendicular to the plane of the paper which is the y-z plane (Fig. 10.31).

Suppose the electric field vector **E** makes an angle θ with the z-axis. We can think of this field **E** as a superposition of two waves—a wave $E \cos \theta$ linearly polarized along the z direction and a wave $E \sin \theta$ linearly polarized along the y-direction. These two waves are in phase, We can now understand the experiment with two polaroids of Fig. 10.30. A polaroid has the property that it transmits that component of the **E** field which is oriented along its axis (shown by a thick arrow with polaroid P_1 in Fig. 10.30) and absorbs the other (perpendicular) component. The basic reason for this action of the polaroid is that it is made from a coloured substance (dye) in which the molecules point along a particular axis when the polaroid is prepared. The movement of the electrons along the axis allows them to absorb the wave polarized in that direction and not the perpendicular component which it transmits almost completely.

In Fig. 10.30, the light transmitted by P_1 is polarized along its axis, this light emerging from P_1 is shown by double arrows. If the axis of the second polaroid P_2 is inclined at an angle θ with that of the first polaroid P_1, then P_2 will transmit a fraction $\cos\theta$ of the incident amplitude; the wave transmitted by P_2 being polarized along *its* axis. The transmitted intensity is a fraction $\cos^2\theta$ of the incident intensity. If P_2 is rotated until its axis is parallel to that of P_1 [as shown in Fig. 10.30(a)], then $\theta = 0°$ and all the intensity incident on P_2 is transmitted by it. But if the axes of P_1 and P_2 are perpendicular to each other [as shown in Fig. 10.30(b)], $\theta = 90°$ and $\cos 90° = 0$ so that no light is transmitted by P_2. The reason is that now the electric field of the wave incident on P_2 is oriented perpendicular to the axis of P_2.

We can now understand the action of a single polaroid placed in the path of ordinary light from, say, a filament bulb. This light is *unpolarized*, i.e. the electric field vector **E** of this light does not have any fixed orientation. The reason is that the radiation from a source like a bulb is produced by millions of electrons in motion which do not emit radiations spontaneously. Hence the orientation of the **E** vector in the transverse plane changes in an irregular fashion with no definite phase relationship between the different components. Light of this kind is called unpolarized. When this light passes through a polaroid, the component parallel to the axis of the polaroid will be transmitted in all orientations of the polaroid because the incident unpolarized light has all the possible orientations in the transverse plane. The component $E\cos\theta$ of each field oscillation will be transmitted, the perpendicular $E\sin\theta$ of each is absorbed. Hence the light emerging from the polaroid is *polarized* along its axis. Such a light is called *linearly polarized*. Refer to supplement S.2 at the end of this chapter for circularly and elliptically polarized lights.

Polarization by Scattering

Look at a clear blue portion of the sky through a polaroid. When the polaroid is rotated, we find that the intensity of light emerging from it shows a rise and fall. This implies that the light incident on the polaroid is polarized. The reason is that the sunlight is scattered (i.e. its direction is changed) when it encounters the molecules of the earth's atmosphere. The scattered light is polarized. Polarization by scattering is explained as follows.

Figure 10.32 shows that the incident light is unpolarized; it has an electric field oscillating in all possible directions in the *y*-*z* plane; the incident wave is travelling in the *x*-direction. This light is incident on one of the molecules of the earth's atmosphere, situated at P, just over the observer O. Owing to the electric field of the incident wave, the positive charges in the molecule experience a force in the direction of the field and the electrons experience a force in the opposite direction. This results in a charge separation in a molecule which therefore becomes an electric dipole. Since the incident field is oscillatory, the dipole also oscillates. Now we know that an oscillating charge itself radiates electromagnetic waves which constitute the scattered light. The two polarizations; one along the *y*-direction (YY′) and the other along the *z*-direction (ZZ′) are shown in the figure. The observer is looking in the *z* direction. The ZZ′ polarization of the scattered light cannot generate an electromagnetic wave along the *z*-direction because electromagnetic waves are transverse (it could generate only a longitudinal wave in the *z*-direction). Hence only the YY′ polarization is possible in the *z*-direction. Thus the wave received by the observer is polarized in the *y*-direction. Similarly the light reaching the observer

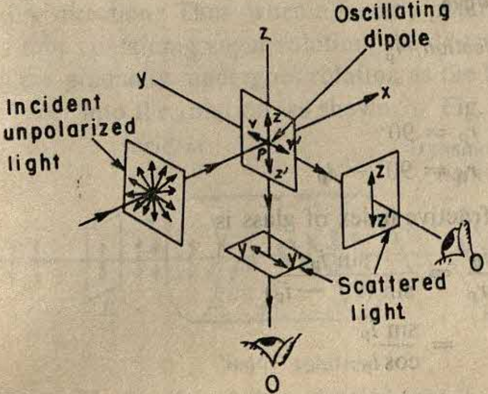


Fig. 10.32 *Polarization of scattered light by an oscillating molecular dipole*

O′ looking along the y-direction will be polarized in the z-direction. Thus we see that the light scattered in any direction in a plane transverse to the direction of the incident light is linearly polarized. In all other directions the scattered light will be only partially polarized.

Polarization by Reflection: Brewster's Angle

In 1808, the French physicist Malus discovered that when a beam of ordinary light is incident at a particular angle on the surface of a transparent medium, the reflected light is polarized.

Figure 10.33 shows an unpolarized light incident at an angle i_p at the surface of glass. Brewster discovered that when $i_p = 57°$, the reflected and refracted rays are exactly 90° apart. The angle i_p when this happens is called the *polarizing* or the *Brewster angle.* When light is incident at the Brewster angle, the reflected wave is linearly polarized. The value of this angle depends upon the refractive index of the medium (glass, in our example) with respect to the medium in which the incident wave is travelling (air, in our example) which in turn depends upon the wavelength of light. If r_p is the corresponding angle of refraction, then from the geometry of Fig. 10.33,

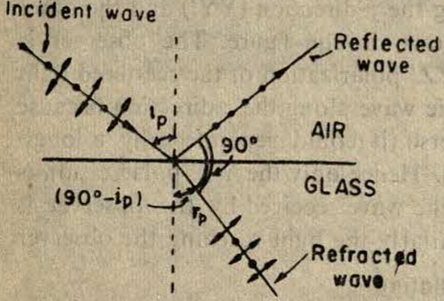


Fig. 10.33 *Polarization by reflection, i_p is Brewster's angle*

$$i_p + r_p = 90°$$

or
$$r_p = 90° - i_p$$

Now from Snell's law, the refractive index of glass is

$$n = \frac{\sin i_p}{\sin r_p} = \frac{\sin i_p}{\sin (90° - i_p)}$$

$$= \frac{\sin i_p}{\cos i_p}$$

$$= \tan i_p$$

$$\therefore \quad n = \tan i_p \qquad (10.26)$$

Equation (10.26) is called the *Brewster law* and the special angle satisfying this equation is the *Brewster angle*. There is a simple explanation for polarization by reflection at the Brewster angle.

The electric field vectors in the unpolarized incident wave can be resolved into two rectangular components, one perpendicular and other parallel to the plane of incidence (the plane of Fig. 10.33). These are shown by dots and double arrows. When the wave reaches the glass surface, the electrons (electric dipoles to be precise) begin to oscillate along these two directions. At the polarizing angle i_p, the reflected and refracted waves are at right angles to each other. It is clear that the oscillations in the plane of incidence are along the direction of the reflected light. These oscillations cannot generate a transverse wave along this direction. Hence the reflected light has field vectors oscillating in a plane perpendicular to the plane of incidence, i.e. the reflected wave is linearly polarized. This is shown by dots. This can be checked by analysing the reflected wave with a polaroid. The refracted ray, however, is not completely polarized.

EXAMPLE 10.13 Find the Brewster angle for air-water interface for yellow light. Given, refractive index of water for yellow light = 1.33.

Solution:

$$\tan i_p = n = 1.33$$

$$\therefore \quad i_p = \tan^{-1}(1.33) = 53^\circ$$

Rotation of Plane of Polarization

When linearly polarized light travels through glass or water, the direction of linear polarization remains fixed. However, many other substances, such as sugar solution, show a remarkable property of rotating this direction. Thus when a linearly polarized light is passed through a tube containing sugar solution, the direction of polarization of the light gradually undergoes rotation as the beam penetrates deeper and deeper into the solution, as shown in Fig. 10.34. Finally it emerges

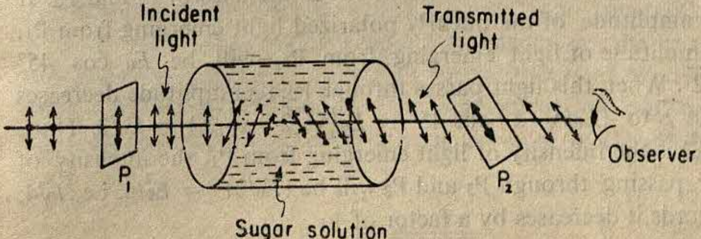


Fig. 10.34 *Rotation of plane of polarization by sugar solution*

with its direction of polarization inclined to that of the incident beam. This phenomenon is called *rotatory polarization* or *optical activity*.

Figure 10.34 schematically illustrates this effect. An unpolarized light is linearly polarized by a polaroid P_1. The light from P_1 is shown passing through a sugar solution in a tube. The direction of polarization rotates as shown. The emergent light is analysed by another polaroid P_2. Before the tube is placed between P_1 and P_2, the polaroids were crossed so that no light emerged from P_2. But when the tube was placed between them, light emerged from P_2, indicating that the direction of polarization of the beam has been rotated. The polaroid P_2 has to be rotated through an angle θ to obtain extinction again. This angle measures the angle through which the direction of polarization has been rotated by the sugar solution. The angle θ is found to be proportional to the length of the solution traversed and the concentration of sugar; it also depends upon the wavelength of light used.

If you notice the polarization vectors in Fig. 10.34 you will find that they rotate like a left-hand screw. The sample of sugar is called *levo-rotatory* or left-handed. Such a substance is not identical to its mirror image which is right-handed or *dextro-rotatory*. Chemists can prepare the right-handed sugar which is a mirror image of the naturally occurring left-handed sample. The right-handed rotation means that the plane of polarization rotates in the clockwise direction with respect to an observer looking towards the oncoming beam of light. A left-handed rotation is in the anticlockwise direction. An equal mixture of the two kinds of sugar solutions shows no optical activity. Optical activity is a molecular phenomenon. Only those molecules which are not identical to their mirror images produce this effect.

EXAMPLE 10.14 Two polaroids P_1 and P_2 are placed at 90° to each other so that the transmitted intensity is zero. What happens when a third polaroid P_3 is placed between them bisecting the angle between them?

Solution: The axis of P_3 is inclined at 45° to the axes of P_1 and P_2. If E_0 is the amplitude of the linearly polarized light emerging from P_1, then the amplitude of light emerging from P_3, will be $E_0 \cos 45^\circ = E_0/\sqrt{2}$. When this light passes through P_2, its amplitude decreases from $E_0/\sqrt{2}$ to $(E_0/\sqrt{2}) \cos 45^\circ = E_0/\sqrt{2} \cdot \sqrt{2}) = E_0/2$. Thus, if $I_0 (= E_0^2)$ is the intensity of light emerging from P_1, the intensity of light after passing through P_3 and P_2 will be $(E_0/2)^2 = E_0^2/4$, i.e. $I_0/4$; in other words,it decreases by a factor of 4.

10. THE DOPPLER EFFECT FOR LIGHT

You have already come across the Doppler effect for sound waves in class XI. You have seen that if a source of sound is moving with a velocity u towards an observer at rest in a medium, the frequency of sound as measured by him is given by

$$\nu' = \nu \left[\frac{1}{1 - (u/v)}\right] \tag{10.27}$$

where v is the speed of sound in the medium and ν is the frequency if the source were at rest.

On the other hand, If the source is at rest and the observer is moving with a velocity u towards the source, the frequency of sound as measured by the approaching observer is given by

$$\nu' = \nu \left(1 + \frac{u}{v}\right) \tag{10.28}$$

Notice that, although the relative speeds u of the source and the observer are the same in the two cases, the frequencies predicted by Eqs (10.27) and (10.28) are different. This is not surprising because (as you have seen in class XI) a source of sound moving through a medium in which the observer is at rest is physically a different situation from that of an observer moving through that medium with the source at rest. We say that the Doppler effect for sound is asymmetrical.

What about the Doppler effect for light? The Doppler effect has been experimentally observed for light as well. We might be tempted to apply Eqs (10.27) and (10.28) to light, substituting c (the speed of light) for v (the speed of sound). But we know that the mechanism of production and propagation of light and sound waves are very different. In the case of light waves travelling in vacuum, a statement that the source or the observer is moving depends upon the frame of reference. In fact, no experiment to date has been able to identify a medium relative to which the source of light or observer are moving. This means that "source moving towards an observer" or "observer moving towards a source" are physically identical situations (as far as light waves are concerned) and must show *the same* Doppler effect. Thus, if applied to light, either Eq. (10.27) or Eq. (10.28) or both must be incorrect. It turns out that both are only approximately correct.

A correct derivation of Doppler frequency in the case of light requires the knowledge of Einstein's special theory of relativity which is outside the scope of this book. For completeness, we only quote (without a derivation) the correct relativistic formula for Doppler frequency:

$$\nu' = \nu \frac{1 + \dfrac{u}{c}}{\sqrt{1 - \dfrac{u^2}{c^2}}} \tag{10.29}$$

Note: Equations (10.27) to (10.29) have been written for the case where the source or observer moves away from the other. The appropriate equations for the case where the source or observer moves towards the other are obtained if we replace u by $-u$.

Notice that the Doppler effect for light is symmetrical, i.e. it does not matter whether the source or the observer or both move, the Doppler frequency remains the same. We will now show that Eqs (10.27) and (10.28) (with v replaced by c) are correct only if the ratio u/c is small enough. Let us expand Eqs (10.27) to (10.29) by the binomial theorem; we respectively get

$$\nu' = \nu\left(1 - \frac{u}{c}\right)^{-1} = \nu\left(1 + \frac{u}{c} + \frac{u^2}{c^2} + \ldots\right)$$

$$\nu' = \nu\left(1 + \frac{u}{c}\right)$$

and
$$\nu' = \nu\left(1 + \frac{u}{c}\right)\left(1 - \frac{u^2}{c^2}\right)^{-1/2}$$

$$= \nu\left(1 + \frac{u}{c} - \frac{1}{2}\frac{u^2}{c^2} + \ldots\right)$$

If u/c is small enough so that the terms of order $(u/c)^2$ can be neglected all the three equations reduce to one equation

$$\nu' = \nu\left(1 + \frac{u}{c}\right) \quad (\text{for } u \ll c) \tag{10.30}$$

In nearly all situations the ratio u/c is small, even for light sources of atomic dimensions. The Doppler shift is $\Delta\nu = (\nu' - \nu)$ and is given by

$$\frac{\Delta\nu}{\nu} = \frac{u}{c} \tag{10.31}$$

for the source and observer moving towards each other in which case $\Delta\nu$ is positive (increase in frequency). When they move away from each other $\Delta\nu$ is negative (decrease in frequency).

The Doppler effect for light finds many applications in astronomy, where it is used to determine the speeds at which the luminous heavenly bodies (stars and galaxies) are moving towards us or away from us by measuring their Doppler shifts. A wavelength in the middle of the visible part of the spectrum will shift towards blue (called the *blue shift*) when the source and observer approach each other and towards red (called the *red shift*) when the source and observer move away or recede from each other. All galaxies for which measurements have been made are observed to show the red shift indicating that they are receding away from us. These observations suggest that the universe is expanding. It must be understood that Doppler shifts measure only the *radial* component (i.e. the line-of-sight component) of the relative velocity u.

EXAMPLE 10.15 The spectral line ($\lambda = 6440$ Å) in the light received from a distant star is observed to be red-shifted by 0.322 Å. Find the speed of the star in the line of sight.

Solution

$$\frac{\Delta\nu}{\nu} = \frac{u}{c}$$

Since $c = \nu\lambda$, we have

$$\frac{\Delta\lambda}{\lambda} = \frac{u}{c}$$

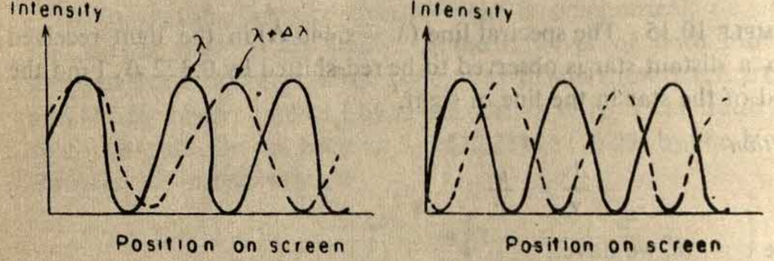
or

$$\begin{aligned} u &= c \cdot \frac{\Delta\lambda}{\lambda} \\ &= 3.0 \times 10^8 \times \frac{0.322}{6440} \\ &= 1.5 \times 10^4 \text{ ms}^{-1} \end{aligned}$$

SUPPLEMENTS

S.1 VISIBILITY OF INTERFERENCE FRINGES

In Young's double slit experiment (Fig. 10.12) we assumed that the light diverging from the source slit S was monochromatic. In fact. no light source is truly monochromatic. Light from every source has a certain wavelength spread $\Delta\lambda$ about a dominant wavelength λ. For example, in the case of sodium light $\Delta\lambda \simeq 6$Å and $\lambda = 5890$Å. Smaller the value of $\Delta\lambda$ compared to λ, the more monochromatic is the source. Nowadays we have lasers which have a very high degree of monochromaticity. Therefore, in an interference experiment, with, for example, sodium light, we get sets of fringes corresponding to the wavelengths present in the sources. These fringes overlap. If $\Delta\lambda$ is such that bright fringes due to one wavelength superpose on the dark fringes due to the other, the entire interference experiment will be washed out. If $\Delta\lambda \ll \lambda$, then in the region close to the central fringe (point P_0), the bright fringes due to different wavelengths will nearly coincide so that the interference pattern is distinctly visible over this small region. Far away this condition is not satisfied leading a poor visibility of fringes [see Fig. 10.35(a)]. It was for this reason that we assumed $y \ll D$.



(a) (b)

Fig. 10.35 *Superposition of interference patterns due to: (a) different wavelengths, and (b) waves from different parts of an extended source*

Another reason for poor visibility of fringes is the finite size of the source slit. Let this size be Δs. Each part of the source produces a fringe pattern on the screen. Since these parts are incoherent, the intensities are simply added. If the maxima of one fall over the minima of the other, the total intensity everywhere on the screen will be the same and no fringes are visible [Fig. 10.35(b)]. Hence the source slit should be very narrow.

S.2 CIRCULARLY AND ELLIPTICALLY POLARIZED LIGHTS

Suppose we have linearly polarized light with electric field **E** oriented at angle $\theta = 45°$ with the x-axis. The amplitude E_0 of the field can be resolved along x and y axes. The x and y components are $E_0 \cos 45°$ and $E_0 \sin 45°$, i.e. $E_0/\sqrt{2}$ and they are phase. Suppose somehow we introduce a phase of $\pi/2$ in one of these components. This can be done by using a thin calcite or quartz plate. If it has the right thickness for a light of a given wavelength, it can introduce the required phase difference between x and y components because it has different refractive indices along x-and y-directions. A plate which introduces a phase difference of $\pi/2$ is called a *quarter wave plate*. Suppose the plate introduces a phase of $\pi/2$ in the x-component only, then we can write

$$E_x = \frac{E_0}{\sqrt{2}} \sin\left(\omega t + \frac{\pi}{2}\right) = \frac{E_0}{\sqrt{2}} \cos \omega t$$

and $$E_y = \frac{E_0}{\sqrt{2}} \sin \omega t$$

Squaring and adding these equations, we get

$$E_x^2 + E_y^2 = E_0^2/2$$

which is the equation of a circle. This means that the tip of the electric field vector **E** describes a circle. Light having this property is referred to as *circularly polarized.* If E_x and E_y do not have equal amplitudes (i.e. if angle θ is not 45°), the quarter wave plate will produce *elliptically polarized* light in which the tip of the electric field vector describes an ellipse.

EXAMPLE 10.16 A half wave plate is a device which introduces a phase difference of π between E_x and E_y. What is its effect on: (a) linearly polarized light with **E** field at an angle θ to the x-axis, and (b) circularly polarized light?

Solution: (a) E_x and E_y are now given by

$$E_x = E_0 \cos\theta \sin(\omega t + \pi) = -E_0 \cos\theta \sin \omega t$$

and $$E_y = E_0 \sin\theta \sin \omega t$$

These equations give

$$\frac{E_y}{E_x} = -\tan\theta$$

i.e. $$E_y = (-\tan\theta)\, E_x$$

which is a straight line inclined at angle $-\theta$ with x-axis.

Hence the linearly polarized light remains linearly polarized but its polarization vector is rotated by an angle $\theta - (-\theta) = 2\theta$.

(b) E_x and E_y of a circularly polarized light, after transmission through the half wave plate become

$$E_x = \frac{E_0}{\sqrt{2}} \cos(\omega t + \pi) = -\frac{E_0}{\sqrt{2}} \cos \omega t$$

and $$E_y = \frac{E_0}{\sqrt{2}} \sin \omega t$$

These equations give $E_x^2 + E_y^2 = E_0^2/2$. Thus the light remains circularly polarized. The negative sign of the amplitude of E_x shows that the sense of rotation of the tip of the **E** vector is now reversed, i.e. it now rotates in the anticlockwise sense.

SUMMARY

Light is an electromagnetic wave in which the electric and magnetic fields vary harmonically in space and time. What propagates is the energy associated with the electric and magnetic fields of the wave. In a vacuum, light travels with a speed of 3×10^8 ms^{-1}.

Like sound waves, light waves exhibit interference and diffraction. Interference effects can be observed only if the interfering light sources are coherent. Sources that have a constant phase relationship are called coherent sources. They must be derived from the same parent source. Laser, fabricated only recently, is a remarkable device that produces coherent, parallel and intense light having a small angular spread. The bending of light around the edges of an obstacle is called diffraction. Interference and diffraction of light can be explained by Huygens' principle.

In ordinary (unpolarized) light, the electric vector has all possible orientations perpendicular to the direction of propagation. In linearly (or plane) polarized light, the electric vector has just one orientation. The fact that light can be polarized suggests that light waves are transverse.

EXERCISES

A. Short-Answer Questions

1. Define a wavefront.
2. The frequency of a wave does not change due to reflection or refraction. Why?
3. What is the geometrical shape of the wavefront in each of the following cases?
   (i) light diverging from a point source, (ii) light emerging out of a convex lens when a point source is placed at its focus, (iii) light emerging out of a concave lens when a parallel beam of light is incident on it, (iv) a parallel beam of light after reflection from a convex mirror, (v) light reflected from a concave mirror when a point source is placed at its focus, and (vi) a portion of the wavefront of light from a distant star intercepted by the earth.
4. The total energy is merely redistributed in the fringe system in any optical experiment. Comment.
5. Why is the central fringe in Lloyd's mirror experiment dark?
6. What is the Brewster angle?

B. Long-Answer Questions

1. State Huygens' principle. Use the principle to show how a parallel beam of light is reflected from a plane mirror. Draw a sketch of the wavefronts and deduce the law of reflection.

2. A parallel beam of light is incident on an air-glass interface. Using Huygens' principle and drawing the sketches of the wavefronts, show how the wave is refracted. Prove the law of refraction.
3. What do you understand by the term "interference" in the context of light? Describe an experimental arrangement to observe it.
4. What are coherent light sources? Why is it not possible to observe interference with noncoherent sources?
5. Draw a diagram of Young's two-slit experiment and show how interference takes place.
6. Deduce the expression for fringe width in Young's two-slit experiment. How will you measure the wavelength of monochromatic light?
7. Describe the production of interference fringes in Lloyd's mirror. In what way do these fringes differ from those produced in Fresnel's prism?
8. Describe and explain how interference bands are produced in a narrow wedge-shaped film. Where are the bands located?
9. Why do we need a broad source to observe interference in thin films?
10. Why are interference effects not observed in thick films?
11. What do you understand by the term "diffraction" in the context of light? Why are diffraction effects in light difficult to observe? Give an example of diffraction of light observed in everyday life.
12. A narrow slit is illuminated by a parallel beam of monochromatic light falling normally on it. Describe what you would observe on a screen placed at the focal plane of a convex lens. Give an elementary explanation of your observation.
13. What do you understand by unpolarized and polarized light? Describe an experiment to illustrate the polarization of light.
14. Describe and explain polarization of light by scattering.
15. What is Brewster's law? Explain polarization of light by reflection. Deduce Brewster's law.
16. Distinguish between circularly and elliptically polarized lights. How are they produced?

C. Multiple-Choice Questions

Choose the correct answer from the given alternatives.

1. How is the interference pattern in Young's double slit experiment affected if sodium light is replaced by violet light of the same intensity?
   (a) The fringes will become brighter.
   (b) The fringes will become fainter.
   (c) The fringe-width will decrease.
   (d) The fringe-width will increase.
2. Monochromatic light from a narrow slit illuminates two narrow parallel slits producing an interference pattern on a screen. The distance between the screen and the slits is now doubled and the distance between the slits is reduced to half. The fringe width
   (a) is doubled (b) becomes four times
   (c) is halved (d) remains unchanged
3. What happens if the monochromatic light used in Young's double-slit experiment is replaced by white light ?

(a) All bright fringes become white.
(b) All bright fringes have colours between violet and red.
(c) Only the central fringe is white, all other fringes are coloured.
(d) No fringes are observed

4. How is the interference pattern in Young's double-slit experiment affected if one of slits, say S_1 is covered with black opaque paper?
   (a) The bright fringes become fainter.
   (b) The fringe width decreases.
   (c) There is uniform illumination all over the screen.
   (d) There is a bright slit but no interference pattern is obtained.
5. What happens if one of slits, say S_1, in the experiment shown in Fig. 10.12 is covered with a glass plate which absorbs half the intensity of light from S_1?
   (a) The bright fringes become less bright and the dark fringes have finite light intensity.
   (b) The bright fringes become brighter and dark fringes become darker,
   (c) The fringe width decreases.
   (d) No fringes will be observed.
6. What happens to the interference pattern if the two slits S_1 and S_2 in Young's double-slit experiment are illuminated by two independent identical sources?
   (a) The intensity of the bright fringes is doubled.
   (b) The intensity of the bright fringes becomes four times.
   (c) Two sets of interference fringes overlap.
   (d) No interference pattern is observed.
7. What is the reason for your answer to Question 6?
   (a) The two sources do not emit light of the same wavelength.
   (b) The two sources emit waves which travel with different speeds.
   (c) The two sources emit light waves of different amplitude.
   (d) There is no constant phase difference between the waves emitted by the two sources.
8. A single-slit diffraction pattern is obtained using a beam of red light. What happens if the red light is replaced by blue light?
   (a) There is no change in the diffraction pattern.
   (b) Diffraction fringes become narrower and crowded together.
   (c) Diffraction fringes becomes broader and farther apart.
   (d) The diffraction pattern disappears.
9. Which one of the following waves cannot be polarized?
   (a) radio waves (b) x-rays
   (c) transverse waves in a string (d) longitudinal waves in a gas
10. The sky looks blue because
    (a) sunlight turns blue after reflection from the planets and the moon
    (b) the sky is illuminated by light from the blue oceans
    (c) sunlight is diffracted by dust particles in the atmosphere
    (d) the light scattered by the dust particles and air molecules is predominantly blue

D. Numerical Problems

1. The index of refraction for a sample of glass for violet light is 1.665 and for red light is 1.618. Calculate the difference in the velocity of light in glass for the two colours.
2. Wavelength of yellow light in air is 6000 Å. Calculate its wavelength in water and glass, the indices of refraction of water and glass being 4/3 and 1.6 respectively.
3. A glass slab of thickness 8 cm contains the same number of waves as 9 cm of water when both are traversed by the same monochromatic light. If refractive index of water is 4/3, calculate the refractive index of glass.
4. Determine the depth of water that contains the same number of waves as 16 cm of glass when each is traversed by the same monochromatic light. The refractive indices of water and glass are 4/3 and 1.5 respectively.
5. In Young's two-slit experiment the separation between 10 fringes is found to be 5.5 mm on a screen placed at a distance of 2.0 m from the slits. If the separation between the slits is 2.00 mm, determine the wavelength of light.
6. What is the distance between two consecutive bright fringes of Problem 5 if the separation between the slits is reduced to 1.0 mm, and the distance of the screen from the slits is reduced to 1.0 m?
7. Using the data of Problem 5, calculate the fringe width for:
    (a) violet light of $\lambda = 4.5\times10^{-7}$ m
    (b) yellow light of $\lambda = 6.0\times10^{-7}$ m
    (c) red light of $\lambda = 7.5\times10^{-7}$ m
8. In Young's double-slit experiment, the intensity of light at a point on the screen, where the path difference is λ, is K units. What is the intensity at a point where the path difference is: (a) $\lambda/2$ and (b) $\lambda/3$?
9. The wedge-shaped film of Fig. 10.18 is illuminated by a light of wavelength 4000 Å falling normally on it. Starting from the point P, 12 dark fringes were counted. Calculate the thickness of the film at the twelfth dark fringe.
10. How many dark fringes would be observed in the same region of the film in Problem 9 if it were illuminated with a light of wavelength 6000 Å.
11. A parallel beam of light of wavelength 6×10^{-7} m is incident normally on a slit of width 0.3 mm. The diffraction pattern is observed on a screen which is placed at the focal plane of a convex lens of focal length 25 cm. Assuming that the lens is placed close to the slit, determine the width of the central maximum of the diffraction pattern.
12. A laser beam of wavelength 7000 Å has an aperture of 1 cm. The beam is focussed on a surface placed at the focal plane of a lens of focal length 5 cm. Compute the areal spread of the beam. Calculate the intensity in watts per square metre if the power of beam is 0.01 W.

11

RAY OPTICS AND OPTICAL INSTRUMENTS

11.1 INTRODUCTION

In the previous chapter we have seen that the concept of rays of light is valid in most common situations in everyday life. The straight-line path of light when it travels from one point to another is called a ray of light. Ray optics, therefore, uses the geometry of straight lines to discuss the reflection of light from spherical mirrors and the refraction of light through prisms and spherical surfaces such as lenses. These simple devices form important components of optical instruments such as cameras, telescopes and microscopes. These aspects of the behaviour of light constitute the subject matter of this chapter.

11.2 PHOTOMETRY

Photometry is a branch of physics that deals with the measurement of light emitted by or reflected from objects. The common sources of light are the sun, electric bulbs, electric arcs and fluorescent tubes. In a narrower sense, photometry is devoted to the measurement of the effect of visible light on the human eye. This effect is characterized by the following photometric quantities: (i) the luminous intensity of the source, (ii) the luminous flux from the source and (iii) the illuminance of a surface that receives light.

When a source emits electromagnetic radiation, the energy is distributed among several wavelengths, visible and invisible to the human eye. In the visible region, the energy is distributed over the wavelength range 390 nm to 760 nm (3900 Å to 7600 Å). The effect of the visible light on the eye depends on the spectral sensitivity of the eye (V_λ) which is defined as the ratio of the luminous flux of the given monochromatic radiation to the energy flux of this radiation (these quantities will be defined a little later). The quantity

$$K_\lambda = \frac{V_\lambda}{(V_\lambda)_{\max}}$$

is called the relative spectral sensitivity of the eye. $K_\lambda = 1$ at $\lambda = 555$ nm. The dependence of K_λ on λ is shown in Fig. 11.1 which is called the standard luminosity curve.

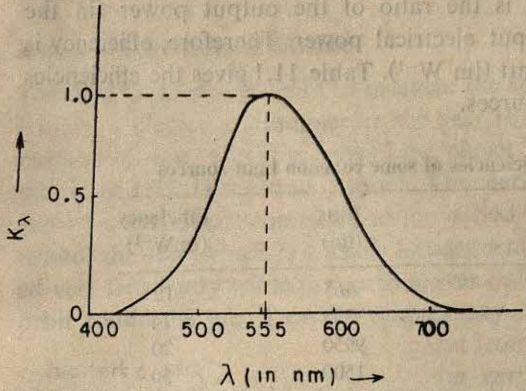


Fig. 11.1 *Standard luminosity curve*

Luminous Intensity (I)

A light source is arbitrarily chosen as a standard (or a reference) and the luminous intensity of any other source is compared with the chosen standard. In the early days, light from a certain type of a candle was used as a standard, and the unit of luminous intensity was called the *candle power*. But now better standards are available. The unit used in the SI system is called *candela* (symbol cd). *The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency* 5.40×10^{14} *Hz and has a radiant intensity af 1/683 watt per steradian in that direction.* (The steradian, symbol sr, is a unit of solid angle and is obtained as follows. Take a sphere of any radius *r*. On the surface of the sphere, cut off an area equal to r^2. The solid angle subtended by this area at the centre of the sphere is equal to one steradian.) The steradian is dimensionless, so that the candela has the dimensions of power (watt) or energy per unit time.

Luminous Flux (Φ)

As stated earlier, a source gives out radiations in a wide range of wavelengths out of which only a narrow band of radiation with wavelengths between 390 and 760 nm is visible to the human eye. *The luminous flux* (*Φ*) *is defined as the energy of the visible part of the radiation emitted, transmitted or received per second.*

Lumen The unit of luminous flux is called lumen (symbol lm) *The* *lumen is the luminous flux emitted within a unit solid angle (steradian) by a uniform point source having luminous intensity of 1 candela.*

We often talk of the efficiency of electric light sources. The efficiency of an electric light source is the ratio of the output power (in the visible region) to the input electrical power. Therefore, efficiency is expressed in lumen per watt (lm W^{-1}). Table 11.1 gives the efficiencies of a few common light sources.

Table 11.1 Efficiencies of some common light sources

Light Source	Flux (lm)	Efficiency (lmW^{-1})
40-W tungsten bulb	465	12
60-W tungsten-bulb	835	14
500-W tungsten lamp	9950	20
30-W fluorescent tube	1500	50

Notice that fluorescent tubes have a fairly high efficiency.

Illuminance (E)

The illuminance of a surface is defined as the ratio of the luminous flux incident on a unit area of the surface. Thus, if A is the area of the surface,

$$E = \frac{\Phi}{A}$$

Lux The unit of illuminance is called *lux*. A lux is the illuminance of 1 lm/m^2.

Illuminance obeys the inverse square law of distance. The illuminance on a surface at a distance r from a point source of light falls off as $1/r^2$, i.e. $E \propto 1/r^2$.

11.3 VELOCITY OF LIGHT

The speed at which light travels in space is one of the most important quantities in physics. The speed of light in free space is the basic foundation on which Albert Einstein developed the special theory of relativity. For many centuries it was believed that the speed of light was infinitely large. Galileo was the first to suggest that the speed of light was large but finite. In 1660 he made an attempt to measure the speed of light. He took a lamp and asked a friend of his to stand at a distance of a kilometer or so in an open field at night. By covering and uncovering the lamp at regular intervals he tried to find how long the light took to reach his friend (at night). Owing to the enormous speed of light, this time was too short to be measured and his experiment failed.

Römer's Astronomical Method

The first successful attempt to measure the speed of light was made by Römer, a Danish astronomer, in the year 1676. He was studying the motion of one of the moons of the planet Jupiter. This moon had a period of 1.77 days round Jupiter. The period of this satellite (i.e. moon) is very small compared to the period at which the earth goes round the sun (which is 1 year). Consequently the satellite was eclipsed very frequently while the earth moves only a small distance on its orbit round the sun. Thus the eclipses may be regarded as signals sent out from Jupiter and observed on the earth just like the covering and uncovering of the lamp in Galileo's experiment except that the distance involved now is astronomical in scale.

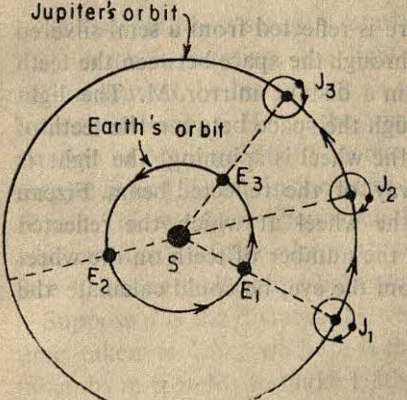


Fig. 11.2 *Römer's method of estimating the speed of light*

Figure 11.2 shows the positions of the earth E_1 and Jupiter J_1 when they are in line with the sun S and closest together. At this position, Römer noted the date and time of the eclipse of one of Jupiter's satellites. After about $6\frac{1}{2}$ months the earth E_2 and Jupiter J_2 are again in line with the sun but farthest apart. Römer noted the time of the eclipse and found that it was $16\frac{1}{2}$ minutes later than the expected time. From this he concluded that this must be the time taken for light to travel a distance equal to the diameter of the earth's orbit round the sun. This distance is about, 3×10^{11} m. So the speed of light should be

$$c = \frac{3 \times 10^{11} \text{ m}}{16.5 \times 60 \text{ s}} \simeq 3 \times 10^{8} \text{ ms}^{-1}$$

After another $6\frac{1}{2}$ months, when the earth is at E_3 and Jupiter at J_3, the eclipse occurred about 16.5 minutes earlier than expected, which is due to the time taken by light to travel the distance = diameter of its orbit round the sun.

Fizeau's Method

In 1849, Fizeau devised the first laboratory method for measuring the speed of light. Figure 11.3 illustrates the principle of his method.

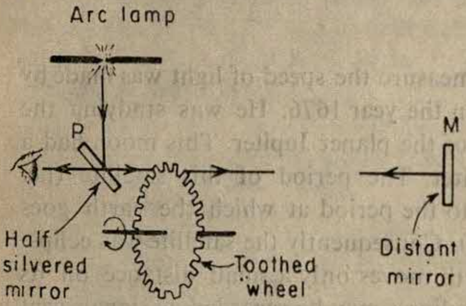


Fig. 11.3 *Fizeau's method*

A beam of light from a carbon arc is reflected from a semi-silvered plate P. The reflected beam passes through the space between the teeth of a toothed-wheel and reflected from a distant mirror M. The light reflected from M again passes through the space between the teeth of the wheel and enters the eye. When the wheel is spinning, the light is cut off if a tooth comes in the way of the reflected beam. Fizeau measured the speed of rotation of the wheel at which the reflected light is cut off. Knowing this speed, the number of teeth on the wheel and the distance of the mirror M from the eye, he could calculate the speed of light.

Michelson's Method

An American physicist AA Michelson spent many yeary of his life in measuring the speed of light. During 1926–29 he devised an accurate method. Figure 11.4 shows a simplified version of his experiment, illustrating the essential principle involved.

M_1 is an equiangular octagonal steel mirror which is mounted on a shaft and can be rotated about a vertical axis through its centre with the help of motor of a variable speed. The faces of the mirror are highly polished. Light from a bright source S passes through a slit and falls at an angle of 45° on a face (say 1) which reflects it to a distant concave mirror M_2. The beam is parallel to the axis of mirror M_2 and strikes it at point A from where it is reflected at a mirror M placed at the focus of M_2. The mirror M reflects the light to B where it is reflected parallel to the axis of the mirror M_2 as shown. If the mirror M_1 is at rest this light will fall on face 3 and will be reflected towards the observer who can see the image of the source through a telescope T.

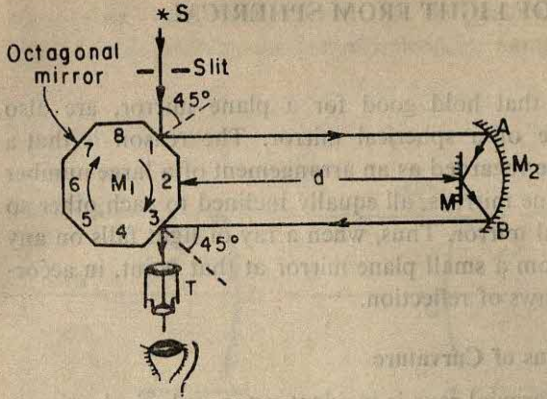


Fig. 11.4 *Michelson's rotating mirror method*

The mirror M_1 is set into rotation starting with a low speed. When the mirror is rotating, the light returning to it from mirror M_2 will not, in general, be incident at an angle of 45° and, therefore, will not enter the telescope. Hence the image disappears. When the speed of rotation of M_1 is gradually increased and suitably adjusted, the image reappears and is seen in the same position as when the mirror M_1 is at rest. This would happen when the speed of rotation is such that the adjacent face 2 of the mirror M_1 is in position formerly occupied by face 1, during which time light travels from M_1 to M_2 and back to M_1.

Suppose d is the distance between mirrors M_1 and M_2. Then the time taken is $2d/c$, where c is the speed of light. But this is the time taken by mirror M_1 to make 1/8th of a revolution which is $1/8n$ where n is the number of revolutions per second. Therefore

$$\frac{1}{8n} = \frac{2d}{c}$$

or

$$c = 16nd$$

The value of n is measured with a stroboscope. Michelson performed this experiment in 1926 and again in 1931 in which the path of light was within an evacuated tube 1.6 km long. Multiple reflections were used to increase the effective path of light and a mirror with 32 faces was also used. His result for the speed of light in vacuum was 2.99774×10^8 ms^{-1}. Michelson died in 1931 while he was working on another more accurate experiment for measuring the speed of light.

Modern methods of measuring the speed of light use highly coherent and unidirectional laser beams which are capable of measuring the speed of light to a very high degree of accuracy. The accepted value of the speed of light in free space is

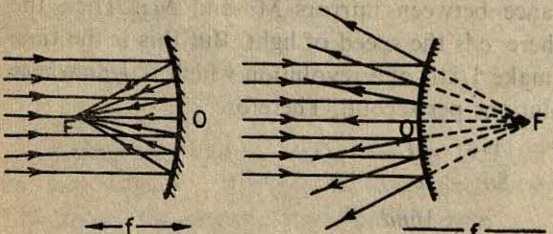
$$c = 2.99792458 \times 10^8 \text{ ms}^{-1}$$

11.4 REFLECTION OF LIGHT FROM SPHERICAL MIRRORS

The laws of reflection that hold good for a plane mirror, are also applicable in the case of a spherical mirror. The reason is that a spherical mirror can be regarded as an arrangement of a large number of extremely small plane mirrors, all equally inclined to each other so as to form a spherical mirror. Thus, when a ray of light falls on any point, it is reflected from a small plane mirror at that point, in accordance with the usual laws of reflection.

Focal Length and Radius of Curvature

If a parallel beam of paraxial rays is incident on a spherical mirror, the reflected rays converge to a point F (in the case of a concave mirror) or appear to diverge from a point F (in the case of a convex mirror). This point F is called the focus (or focal point) of the spherical mirror (see Fig. 11.5). Paraxial rays are those that subtend a small angle with the normal to the surface at the point where they strike. The rays are paraxial if they are close to the principal axis of the mirror or if the mirror has a small aperture. The distance OF ($=f$) is called the focal length of the mirror. The distance of the centre of curvature of the mirror from any point on its surface, is the radius of curvature (R) of the mirror.



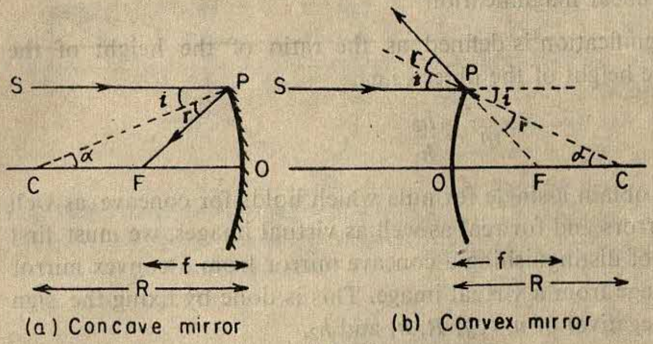
(a) Concave mirror

(b) Convex mirror

Fig. 11.5 *Reflection from concave and convex mirrors*

Relation Between R and f If we restrict ourselves only to the paraxial rays, the relation between R and f is very simple. To obtain this relation, consider a ray of light SP parallel to the principal axis, incident on a spherical mirror (concave or convex) at the point P. The normal to the surface at the point P is CP and CO = CP = R is the radius of curvature. After reflection from the mirror, the ray converges to the point F [in the case of a concave mirror, Fig. 11.6(a)] or appears to diverge from the point F [in the case of a convex mirror, Fig. 11.6(b)] obeying the law of reflection, namely,

$$i = r$$



(a) Concave mirror (b) Convex mirror

Fig. 11.6 *Relation between R and f*

Now since SP is parallel to OC

$$\alpha = i$$

In ΔCPF, $\alpha = r$

Hence $\text{PF} = \text{FC}$ $(\because \quad i = r)$

For paraxial rays, i.e. P very close to O

$$\text{PF} = \text{OF}$$

or $\text{FC} = \text{OF}$

or $\text{OC} = \text{FC} + \text{OF} = \text{OF} + \text{OF} = 2(\text{OF})$

or $R = 2f \quad \text{or} \quad f = R/2$

Thus *the focal length of a spherical mirror is half its radius of curvature.* Remember that this result is true *only for paraxial rays.*

The Spherical Mirror Formula

We shall now derive a mathematical relation between the object and image distances and the focal length of a spherical mirror. This relation is called the spherical mirror formula. We shall see that a single formula is applicable to both types of images (real and virtual) and for both types of spherical mirrors (concave and convex).

The symbols used are:

u = the object distance

v = the image distance

f = the focal length

R = the radius of curvature
h_1 = the height of the object
h_2 = the height of the image
m = the linear magnification

Linear magnification is defined as the ratio of the height of the image and the height of the object, i.e.

$$m = \frac{h_2}{h_1}$$

If we wish to obtain a single formula which holds for concave as well as convex mirrors and for real as well as virtual images, we must first find a means of distinguishing a concave mirror from a convex mirror and a real image from a virtual image. This is done by fixing the sign (positive or negative) of u, v, f, R, h_1 and h_2.

The Sign Convention Different writers use different sign conventions, but each convention serves the purpose as well as the other. In this chapter and the next we shall use the so-called "New Cartesian Convention" which is based upon the conventions of coordinate geometry as shown in Fig. 11.7. The pole O of the mirror is taken to be at the origin of the rectangular coordinate system. Accordingly, we shall use the following convention regarding the signs of various quantities.

1. Distances are measured from the pole O of the mirror.
2. Distances measured in the direction of the incident ray are positive.
3. Distances measured in a direction opposite to the incident ray are negative.
4. Heights measured upwards and perpendicular to the x-axis are positive.

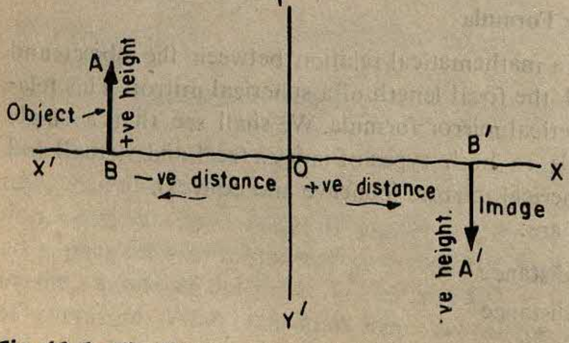


Fig. 11.7 *The Cartesian sign convention*

5. Heights measured downwards and perpendicular to the x-axis are negative.

According to our sign convention, the focal length and the radius of curvature are negative for a concave mirror and positive for a convex mirror. The magnification is negative for a real image and positive for a virtual image.

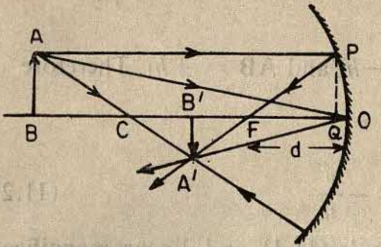


Fig. 11.8 *Relation between u, v, and f for a concave mirror forming a real image*

Consider the ray diagram shown in Fig. 11.8. A′B′ is the image of an object AB formed by reflection from a concave mirror. Since triangles ABC and A′CB′ are similar, we have

$$\frac{AB}{A'B'} = \frac{BC}{B'C}$$

Also from similar triangles PFQ and B′FA′, we have

$$\frac{PQ}{A'B'} = \frac{QF}{FB'}$$

Since PQ = AB,

$$\frac{BC}{B'C} = \frac{QF}{FB'}$$

For paraxial rays, i.e. P very close to O (valid for small apertures), Q will be very close to O so that $QF \simeq OF$. Thus we have

$$\frac{BC}{B'C} = \frac{OF}{FB'}$$

or

$$\frac{OB - OC}{OC - OB'} = \frac{OF}{OB' - OF}$$

or

$$\frac{OB - 2OF}{2OF - OB'} = \frac{OF}{OB' - OF} \qquad (\because\ OC = 2OF)$$

Now, from our sign convention, $OB = -u$, $OB' = -v$, and $OF = -f$. Therefore, we have

$$\frac{-u - (-2f)}{-2f - (-v)} = \frac{-f}{-v - (-f)}$$

which gives

$$\frac{1}{v} + \frac{1}{u} = \frac{1}{f} \qquad (11.1)$$

Linear Magnification Referring to Fig. (11.8), we have, since triangles AOB and A′OB′ are similar,

$$\frac{A'B'}{AB} = \frac{OB'}{OB}$$

From our sign convention, A′B′ $= -h_2$ and AB $= +h_1$. Therefore

$$\frac{-h_2}{+h_1} = \frac{-v}{-u}$$

or
$$m = \frac{h_2}{h_1} = -\frac{v}{u} \tag{11.2}$$

We have obtained the mirror formula (11.1) and linear magnification (11.2) for a concave mirror forming a real image. It is easy to see that these formulae are applicable to a concave mirror forming a virtual image and to a convex mirror which forms only virtual images.

An Important Note Students often experience difficulties in the use of the sign convention while solving numerical problems. To solve any problem on a spherical mirror, remember the following procedure:

1. Use the mirror formula $\frac{1}{v} + \frac{1}{u} = \frac{1}{f}$ for concave as well as convex mirrors.
2. Substitute the numerical values of the given quantities with the proper positive or negative sign.
3. Do not give any sign to the quantity to be determined. The unknown quantity will be obtained with its proper sign.
4. Remember that linear magnification $m = h_2/h_1$ is *negative* for a *real* image and *positive* for a *virtual* image. This follows from the fact that a real image is inverted and a virtual image erect.

It is always useful to draw a rough figure when you start solving a numerical problem. The following solved examples illustrate the use of the sign convention.

EXAMPLE 11.1 A concave mirror forms a real image four times as tall as the object placed 10 cm in front of the mirror. Find the position of the image and the radius of curvature of the mirror.

Solution: Since the image is real, the magnification is negative i.e.

$$m = -4$$

$$u = -10 \text{ cm}$$

Now $m = -v/u$. Substituting for m and u, we have

$$-4 = -\frac{v}{-10}$$

or $$v = -40 \text{ cm}$$

The −ve sign here is expected since the image is real. Substituti for u and v in $\frac{1}{v} + \frac{1}{u} = \frac{1}{f}$, we get

$$\frac{1}{f} = \frac{1}{-40} + \frac{1}{-10} = -\frac{1}{8}$$

or $$f = -8 \text{ cm}$$

∴ The radius of curvature $= 2f = -2 \times 8 = -16$ cm

The −ve sign here also is expected since the mirror is concav Thus the image is formed at a distance of 40 cm from the mirror, c the same side as the object and the radius of curvature of the mirr is 16 cm.

Example 11.2 A convex mirror of radius of curvature 20 cm forn an image which is half the size of the object. Locate the position c the object and its image.

Solution: Since the convex mirror can form only virtual image, i magnification must be positive,

$$m = +\frac{1}{2}$$

but $$m = -\frac{v}{u}$$

∴ $$v = -\frac{u}{2}$$

Substituting this value of v in $\frac{1}{v} + \frac{1}{u} = \frac{1}{f}$, we have

$$\frac{1}{-u/2} + \frac{1}{u} = \frac{1}{f} = \frac{1}{10}$$

or $$u = -10 \text{ cm}$$

$$v = -\frac{u}{2} = +5 \text{ cm}$$

Thus the object must be placed at a distance of 10 cm in front o the mirror and its image is formed 5 cm behind the mirror.

Example 11.3 A square wire of side 3.0 cm is placed 25 cm awa from a concave mirror of focal length 10 cm. What is the area enclosed by the image of the wire? The centre of the square is on the axis o the mirror, with its two sides normal to the axis.

Solution: $u = -25$ cm, $f = -10$ cm. The distance of the image is given by

$$\frac{1}{v} = \frac{1}{f} - \frac{1}{u} = \frac{1}{-10} - \frac{1}{-25}$$

or $$v = -\frac{50}{3}\text{ cm}$$

Area of the object wire $= 3.0 \times 3.0 = 9.0\text{ cm}^2$. The area magnification is given by

$$\frac{\text{Area of image}}{\text{Area of object}} = (\text{linear magnification})^2 = \frac{v^2}{u^2}$$

$$= \left(\frac{50}{3 \times 25}\right)^2 = \frac{4}{9}$$

$$\therefore \quad \text{Area enclosed by image} = \frac{4}{9} \times 9\text{ cm}^2 = 4\text{ cm}^2$$

Example 11.4 A boy 1.50 m tall with his eye level at 1.38 m stands before a mirror fixed on a wall. Indicate by means of a ray diagram how the mirror should be positioned so that he can view himself fully. What should be the minimum length of the mirror? Does the answer depend on the eye level?

Solution: The ray diagram is shown in Fig. 11.9 where AB represents the boy with A as his head and B as his feet (AB = 1.5 m). E represents the position of his eyes so that PB = 1.38 m. He will see his full-length image in the mirror MN if the rays which start from A and B reach his eyes after reflection from the mirror. Let P and Q be the mid-points of AE and EB respectively. PM and QN are perpendicular to the mirror. The rays AM and BN will be reflected along ME and NE respectively. These rays enter his eyes, thus enabling him to see his full image A′B′. Thus MN is the shortest length of the mirror. Now

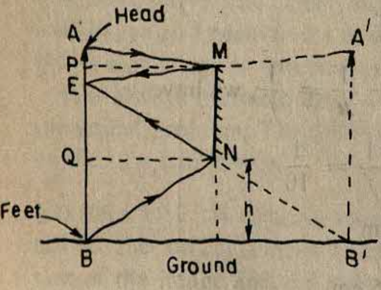


Fig. 11.9

$$MN = PQ = PE + EQ$$

$$= \frac{1}{2}AE + \frac{1}{2}EB$$

$$= \frac{1}{2}(\text{AE} + \text{EB})$$

$$= \frac{1}{2}\,\text{AB}$$

$$= \frac{1.5 \text{ m}}{2} = 0.75 \text{ m}$$

This is the minimum length of the mirror he should use and this does not depend on the eye level, but the way the mirror must be positioned on the wall will depend his eye level.

For the foot to be seen the bottom edge N of the mirror should not be higher that a height h from the ground (see figure) where h is given by

$$h = \text{QB} = \text{EB} - \text{EQ}$$

$$= \text{EB} - \frac{1}{2}\,\text{EB} = \frac{1}{2}\,\text{EB} = \frac{1}{2} \times 1.38 = 0.69 \text{ m}$$

For the head to be seen by the eyes, the top edge M of the mirror should not be lower than MN + h = 0.75 + 0.69 = 1.44 m from the ground.

EXAMPLE 11.5 A small strip of plane mirror A is set with its plane normal to the principal axis of a convex mirror B and placed 15 cm in front of B which it partly covers. An object is placed 30 cm from A and the two virtual images formed by reflection in A and B coincide without parallax. Find the radius of curvature of B.

Solution:

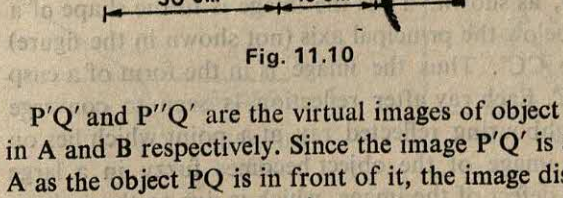


Fig. 11.10

P′Q′ and P″Q′ are the virtual images of object PQ due to reflections in A and B respectively. Since the image P′Q′ is as far behind mirror A as the object PQ is in front of it, the image distance for the convex mirror B is (see Fig. 11.10).

$$v = +\text{OQ}' = +15 \text{ cm}$$

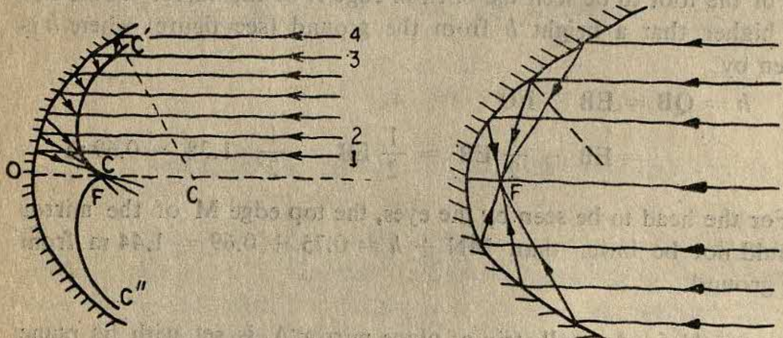
$$\text{Object distance} \quad u = -\text{OQ} = -45 \text{ cm}$$

Using these values in $\frac{1}{v} + \frac{1}{u} = \frac{1}{f}$ gives $f = 22.5$ cm.

Hence $$R = 2f = 45 \text{ cm}$$

Spherical Aberration

In deducing the spherical mirror formula, we assumed that the incident and reflected rays subtend very small angles with the principal axis of the mirror. This assumption is not justified if the mirror has a large aperture.



(a) (b)

Fig. 11.11 *(a) Spherical aberration in a spherical mirror (b)N o aberration in a parabolic mirror*

Figure 11.11(a) shows a parallel beam of light reflected from a concave mirror of a large aperture. Each reflected ray is drawn by using the law of reflection at each point of incidence on the surface of the mirror. We find that only the paraxial rays such as 1 and 2, which are close to the principal axis are focussed at F. The marginal rays, such as 3 and 4, which are near the periphery (or margin) of the mirror are not focussed at F; they strike the principal axis at points closer to the pole O, as shown. Thus the image is in the shape of a curve CC′. The rays below the principal axis (not shown in the figure) form a similar curve CC″. Thus the image is in the form of a cusp C′CC″ with apex at C. Each ray after reflection is seen to converge and intersect its neighbouring reflected ray at a point which lies on the cusp. Hence the image of the object becomes fuzzy in a large aperture mirror. This defect of the image, which is due to the spherical nature of the reflecting surface is called *spherical aberration.*

This failure of rays to converge to or diverge from a single point after reflection from a spherical mirror is called spherical aberration. This defect destroys the sharpness of the image. Spherical aberration is considerably reduced by having a mirror of a small aperture or by putting a stop in front of marginal rays. Since this defect is due to the spherical nature of the reflecting surface, it should be possible to suitably adjust the curvature of the surface in such a way that both paraxial and marginal rays are focussed at a single point. A parabolic mirror shown in Fig. 11.11(b) has a surface that satisfies this requirement. In a parabolic mirror, the surface is that of a paraboloid; a three-dimensional surface obtained by rotating a parabola about its major axis. Parabolic mirrors do not show spherical aberration. It is for this reason that mirrors used in searchlights or headlights in automobiles are parabolic.

11.5 REFRACTION OF LIGHT

In the previous chapter we deduced the law of refraction which states as follows; when a ray of light travels from one medium to another, then for all angles of incidence, the ratio of the sine of the angle of incidence and the sine of the angle of refraction is a constant for the given pair of media; the constant is called the refractive index of the second medium with respect to the first, i.e.

$$n = \frac{\sin i}{\sin r} \tag{11.3}$$

This exact relationship between the angles of incidence and refraction was discovered in 1621 by a Dutch scientist, Willibord Snell and is called the *Snell's law* (Fig. 11.12).

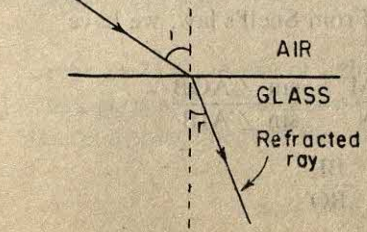


Fig. 11.12 *Refraction of light*

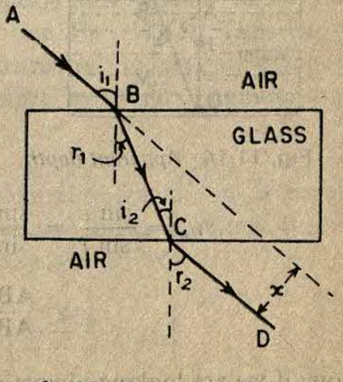


Fig. 11.13 *Refraction through a parallel sided slab*

When a ray of light travels from one medium to a second medium and then to a third medium, refraction occurs at both the interfaces. Figure 11.13 shows a ray AB incident on a parallel-sided glass slab and suffering refractions at B and C. It is easy to see that the emergent ray CD is parallel to the incident ray AB. Since the medium on both sides of the glass slab is the same, we have

$$\frac{\sin i_1}{\sin r_1} = n \qquad \text{and} \qquad \frac{\sin i_2}{\sin r_2} = \frac{1}{n}$$

But $i_2 = r_1$. Therefore,

$$\frac{\sin i_1}{\sin r_1} = \frac{\sin r_2}{\sin r_1}$$

or

$$\sin i_1 = \sin r_2$$

or

$$i_1 = r_2$$

i.e. the emergent ray is parallel to the incident ray. Notice that the emergent ray is displaced with respect to the incident ray. The displacement x depends upon the angle of incident i, refractive index n and the thickness t of the slab (see Example 11.6).

Some Effects of Refraction

1. The apparent depth of objects is less in denser media than the real depth. Figure 11.14 shows an object O lying at the bottom of a beaker containing water. Two rays OA and OB from O are refracted along AD and BC respectively. These divergent rays appear to come from I which is the virtual image of O which is at a smaller depth than O.

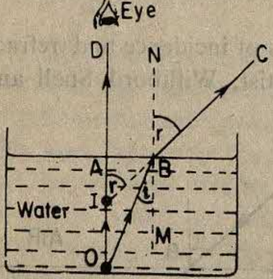


Fig. 11.14 *Apparent depth*

From Snell's law, we have

$$ {}^{w}n_{a} = \frac{\sin i}{\sin r} = \frac{\sin \angle OBM}{\sin \angle CBN} = \frac{\sin \angle AOB}{\sin \angle AIB}$$

$$= \frac{AB/BO}{AB/BI} = \frac{BI}{BO}$$

Now, if we are looking almost vertically downwards, the ray BC will not enter our eye because of the small size of the pupil. Hence, in order to see the image, point B should be close to A, in which case, BI $\simeq$ AI and BO $\simeq$ AO, so that

$$^{a}n_{w} = \frac{AO}{AI} = \frac{\text{real depth}}{\text{apparent depth}}$$

This explains why a swimming pool appears shallower than it really is.

2. The sun is visible before actual sunrise and after actual sunset because of atmospheric refraction.

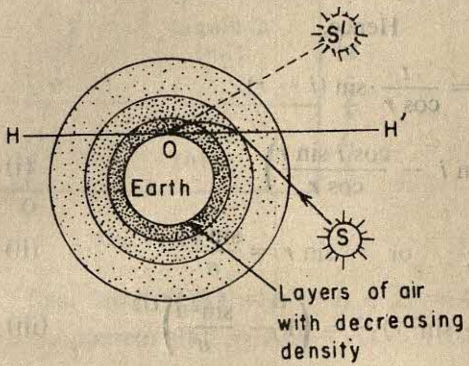


Fig. 11.15 *Refraction effects at sunrise and sunset*

As we go higher up, the density and refractive index of the layers of the atmosphere decrease. The rays starting from the sun pass from the rarer to the denser medium and keep bending towards the normal until they reach the eye of the observer on earth.

Imagine an observer at O with HH′ as the horizon. When the sun S is below the horizon, it actually appears at S′ due to refraction in different layers of the atmosphere (see Fig. 11.15). Thus the sun appears to rise early and (for the same reason) to set late by about 2 minutes. The day thus becomes longer by about 4 minutes due to refraction of light.

EXAMPLE 11.6 A ray of light is incident at an angle i on a parallel-sided glass slab of thickness t and refractive index n. Show that the lateral displacement of the emergent ray is given by

$$x = t \sin i \left[1 - \frac{\cos i}{(n^2 - \sin^2 i)^{1/2}}\right]$$

Can x exceed t?

Solution: Figure 11.16 shows the path of the rays through the slab.

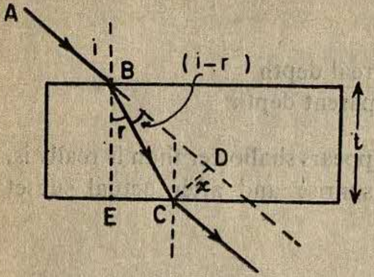


Fig. 11.16

In triangle BCD,

$$\frac{x}{\text{BC}} = \sin (i - r)$$

or $$x = \text{BC} \sin (i - r)$$

In triangle BCE, we have

$$\text{BC} = \frac{\text{BE}}{\cos r} = \frac{t}{\cos r}$$

Hence,

$$x = \frac{t}{\cos r} \cdot \sin (i - r)$$

or $$x = t\left(\sin i - \frac{\cos i \sin r}{\cos r}\right) \tag{i}$$

Now $$n = \frac{\sin i}{\sin r} \quad \text{or} \quad \sin r = \frac{\sin i}{n} \tag{ii}$$

and $$\cos r = (1 - \sin^2 r)^{1/2} = \left(1 - \frac{\sin^2 i}{n^2}\right)^{1/2} \tag{iii}$$

Using (ii) and (iii) in (i) and simplifying, we get

$$x = t \sin i\left[1 - \frac{\cos i}{(n^2 - \sin^2 i)^{1/2}}\right]$$

It is clear that the displacement x tends to a maximum value equal to t if $i \rightarrow 90°$, so that $\sin i \rightarrow 1$ and $\cos i \rightarrow 0$. Hence x cannot exceed t.

EXAMPLE 11.7 The bottom of a container is a 4.0 cm thick glass ($n = 1.5$) slab. The container contains two immiscible liquids A and B of depths 6.0 cm and 8.0 cm respectively. What is the apparent position of a scratch on the outer surface of the bottom of the glass slab when viewed through the container? Refractive indices of A and B are 1.4 and 1.3 respectively.

Solution: Referring to Fig. 11.17, the apparent depth of the scratch is

$$\begin{aligned}\text{AI} &= \frac{t_1}{n_1} + \frac{t_2}{n_2} + \frac{t_3}{n_3} \\ &= \frac{4.0}{1.5} + \frac{6.0}{1.4} + \frac{8.0}{1.3} \\ &= 13.1 \text{ cm}\end{aligned}$$

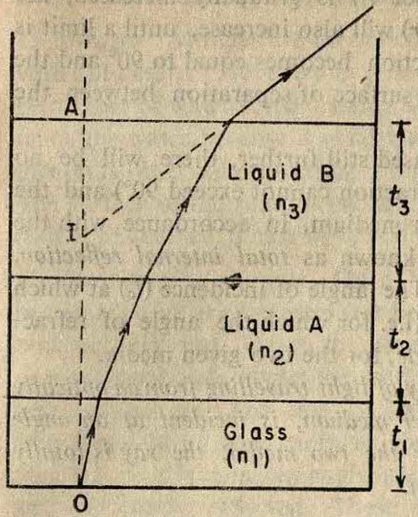


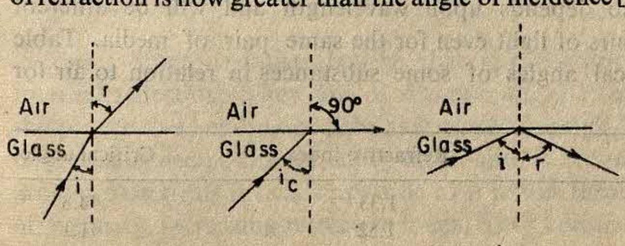
Fig. 11.17

$$\text{Real depth (OA)} = t_1 + t_2 + t_3 = 4.0 + 6.0 + 8.0 = 18.0 \text{ cm}$$
$$\text{Apparent shift} = \text{AO} - \text{AI} = 18.0 - 13.1 = 4.9 \text{ cm}$$

11.6 TOTAL INTERNAL REFLECTION: CRITICAL ANGLE

When a ray of light passes from an optically rarer medium into an optically denser medium, the refracted ray is bent towards the normal so that even when the angle of incidence has a maximum value of 90°, a refracted ray is possible (since the corresponding angle of refraction is less than 90°).

On the other hand, when a ray of light passes from an optically denser medium (such as glass) into an optically rarer medium (such as air), the refracted ray is bent away from the normal, so that the angle of refraction is now greater than the angle of incidence [see Fig. 11.18(a)]



(a) $i < i_c$ (b) $i = i_c$ (c) $i > i_c$

Fig. 11.18 *Total internal reflection and critical angle*

Therefore, if the angle of incidence (i) is gradually increased, the corresponding angle of refraction (*r*) will also increase, until a limit is reached when the angle of refraction becomes equal to 90° and the refracted ray travels parallel to the surface of separation between the two media [see Fig. 11.18(b)].

If the angle of incidence is increased still further, there will be no refracted ray (since the angle of refraction cannot exceed 90°) and the ray is reflected back into the denser medium, in accordance with the ordinary laws of reflection. This is known as *total internal reflection*, and is shown in Fig. 11.18(c). The angle of incidence (i_c) at which *total internal reflection* just occurs (i.e. for which the angle of refraction is 90°) is called the *critical angle*, for the two given media.

Thus we conclude that *when a ray of light travelling from an optically denser medium into an optically rarer medium, is incident at an angle greater than the critical angle for the two media, the ray is totally reflected back into the same medium.*

Relation between Refractive Index and Critical Angle

According to Snell's law, the refractive index of glass with respect to air is given by

$$\frac{1}{{}^{a}n_{g}} = \frac{\sin i}{\sin r}$$

where i = angle of incidence in the denser medium (glass) and r = angle of refraction in the rarer medium (air). When $r = 90°$, $i = i_c$, so that

$$\frac{1}{{}^{a}n_{g}} = \frac{\sin i_c}{\sin 90°} = \sin i_c$$

or

$${}^{a}n_{g} = \frac{1}{\sin i_c} \qquad (11.4)$$

Since refractive index depends upon the wavelength of light used, the critical angle also depends upon wavelength and will be different for different colours of light even for the same pair of media. Table 11.2 shows critical angles of some substances in relation to air for $\lambda = 590$ nm.

Substance	Refractive index	Critical angle
Water	1.33	48.8°
Crown glass	1.52	41.1°
Dense flint glass	1.65	37.3°
Diamond	2.42	24.4°

Some Effects and Applications of Total Internal Reflection

1. *Mirages* Mirages are an optical illusion and are usually associated with hot deserts. Travellers in a hot desert often see what appears to be a sheet of water some distance ahead of them. They are never able to reach the water because it is nothing but hot sand. The same phenomenon is observed on a straight tar road on a hot summer afternoon. The surface of the road, some distance ahead, appears wet and shiny as if it has rained. These optical illusions are called *mirages.*

Explanation of a mirage: On a hot summer day, the layers of air in contact with the tar road (tar being black absorbs heat very quickly) become very hot; hence the densities and refractive indices of the layers immediately above the road are lower than those of the cooler layers higher up. The rays of light from a distant object (such as a tree) bend more and more as they pass through these layers, until they fall on a layer at an angle greater than the critical angle for it and get totally reflected. These reflected rays then travel up and undergo a series of refractions, but in a direction opposite to those in the first case, because they now pass from layers of increasing refractive indices. These rays finally reach the observer's eyes, who sees an inverted image of the object, as though reflected from a pool of water (Fig. 11.19).

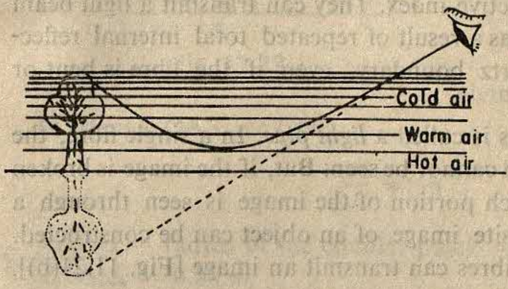


Fig. 11.19 *Formation of a mirage*

Note: Some scientists believe that mirages are not due to total internal reflection. They are of the opinion that mirages are due to the progressive bending of light as it passes through the warmer layers of gradually decreasing refractive index until it becomes parallel to the ground. After this it bends upwards as it passes through cooler layers of gradually increasing refractive index. Thus continuous and progressive refraction (rather than total internal reflection) is believed to be responsible for the formation of a mirage.

2. *Total Reflecting Prisms* Total reflecting prisms are usually made of crown glass of refractive index 1.5. They are right-angled prisms with angles 45°, 45° and 90°. These prisms are used to turn a beam of light through 90° or 180° without any loss of intensity and multiple reflections as in the case of plane mirrors which can also perform this function. The critical angle for total internal reflection for crown glass ($n = 1.5$) is about 42°. Figure 11.20 shows how this prism turns a ray through 90° or 180°. Total reflecting prisms are used in binoculars and submarine periscopes.

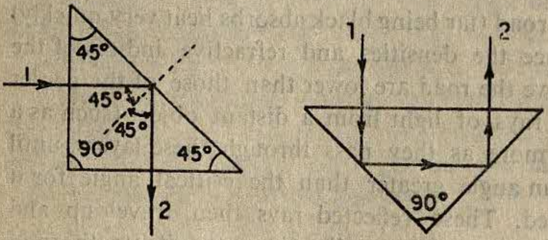


Fig. 11.20 *Total reflecting prisms*

3. *Optical Fibres* Optical fibres are based on the principle of total internal reflection. The fibres are extremely thin (diameter ~ 0.0002 cm) and long strands of glass or quartz. They are coated with a thin layer of a material of lower refractive index. They can transmit a light beam from one end to the other as a result of repeated total internal reflections at the glass or quartz boundary, even if the fibre is bent or twisted [Fig. 11.21(a)].

A bundle of optical fibres is called a *light pipe*. In a single fibre, the complete image of an object cannot be seen. But, if the image is broken into a series of dots and each portion of the image is seen through a separate fibre, the composite image of an object can be constructed. Thus, a bundle of optical fibres can transmit an image [Fig. 11.21(b)].

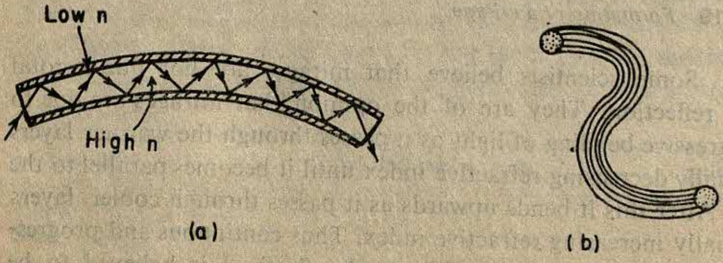


Fig. 11.21 *(a) An optical fibre* *(b) A light pipe*

The advantage of such a pipe is that it is flexible and can be twisted in any desired manner. Therefore, it finds applications in medicine and optical experiments. With the help of this device it is possible to examine the inaccessible parts of an equipment or of the human body (like the interior of lungs and stomach). Optical fibres are also used for transmitting and receiving electrical signals which are converted to light by transducers. They are used in telephone and other transmitting cables. Each fibre can carry as many as 2000 telephone conversations with much less intensity loss.

EXAMPLE 11.8 A small bulb is placed at the bottom of a tank containing water to a depth of 80 cm. What is the area of the surface of water through which light from the bulb can emerge out? Refractive index of water is 1.33. (Consider the bulb to be a point source.)

Solution: It is clear from Fig. 11.22 that the light from the point source S can emerge out of a circular patch of diameter AB. Rays of light from S which are incident at an angle $i > i_c$, i.e. beyond A and B, are totally reflected within water and cannot emerge out of the surface. The critical angle i_c is given by

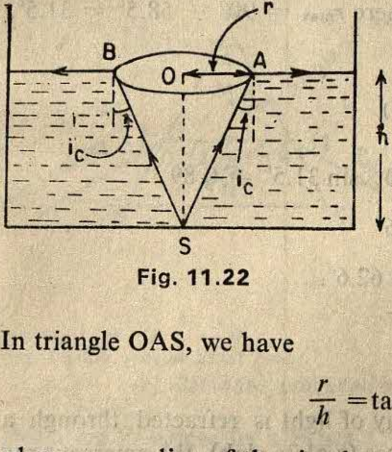


Fig. 11.22

$$\sin i_c = \frac{1}{n} = \frac{1}{1.33} = 0.75$$

giving $i_c = 48.75°$

In triangle OAS, we have

$$\frac{r}{h} = \tan i_c$$

where r = radius of the circular patch = OA or OB and h = depth of water = 0.80 m

$$r = h \tan i_c$$

$$= 0.80 \tan 48.75° = 0.912 \text{ m}$$

$$\text{Area of patch} = \pi r^2 = 3.142 \times (0.912)^2 = 2.6 \text{ m}^2$$

EXAMPLE 11.9 Figure 11.23 shows a cross-section of a thin and long light pipe made of a glass fibre of refractive index 1.70. The outer covering is made of a material of refractive index 1.45. Find the range of angles of incidence (i) for which total reflections can occur inside the pipe.

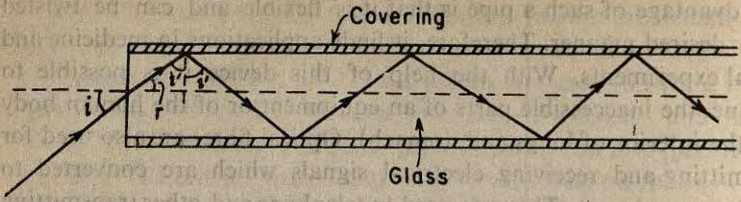


Fig. 11.23

Solution: The critical angle i'_c is given by

$$\sin i'_c = \frac{1.45}{1.70} = 0.853$$

which gives $$i'_c = 58.5^\circ$$

Total internal reflection will occur if the angle $i' > i'_c$, i.e. if $i' > 58.5^\circ$ which implies that angle $r < r_{\max}$ where $r_{\max} = 90^\circ - 58.5^\circ = 31.5^\circ$. Now, from Snell's law

$$\frac{\sin i_{\max}}{\sin r_{\max}} = 1.70$$

$$\therefore \quad \sin i_{\max} = 1.70 \times \sin 31.5^\circ = 0.89$$

or $$i_{\max} = 62.6^\circ$$

i.e. angle i must lie in the range 0 to 62.6°.

11.7 REFRACTION IN A PRISM

We have already seen that when a ray of light is refracted through a transparent medium with parallel faces (a glass slab), the emergent ray is displaced parallel to the incident ray. We shall now examine what happens when the faces are not parallel but are inclined to each other at an angle. A portion of a transparent medium bounded by two plane faces, inclined to each other at an angle, is called a *prism* [Fig. 11.24(a)]. The angle between the two refracting faces DEBA and DACF is called the *refracting angle* (or simply the *angle*) of the prism. The line AD where the two refracting faces meet is called the *refracting edge* of the prism. Any section of the prism made by a plane at right angles to the refracting edge is called the *principal section* of the prism. It is evident that the principal section of the prism in the plane of the paper is ABC with A as the angle of the prism.

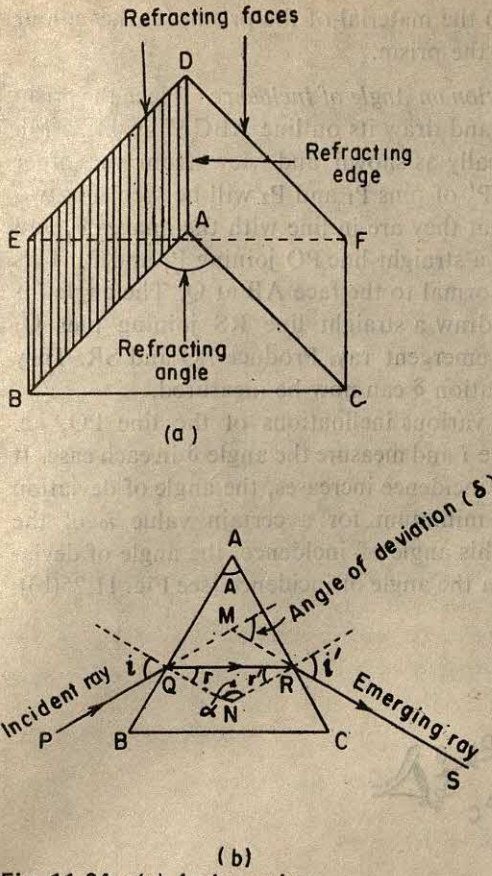


Fig. 11.24 *(a) A glass prism*
(b) Deviation produced by a prism

Deviation Produced by a Prism

A prism does not displace but deviates an incident ray. The angle of deviation is the angle between the incident and emergent rays. Figure 11.24(b) illustrates the action of a prism. ABC is the principal section of a triangular prism shown in Fig. 11.24(a) in the plane of the paper. PQ is a ray of monochromatic light (i.e. light of a single colour) incident on face AB of the prism. This ray enters a denser medium and therefore bends towards the normal and travels along QR in the prism. At face AC, the ray QR suffers another refraction, bending away from the normal as it emerges along RS. The angle of deviation is the angle between the incident and the emergent rays produced as shown.

The value of the angle of deviation depends upon four parameters:

(a) the angle of incidence, (b) the material of the prism, (c) the colour of light, and (d) the angle of the prism.

Dependence of Angle of Deviation on Angle of Incidence Place the prism on a white sheet of paper and draw its outline ABC [Fig. 11.25(a)]. Fix two pins P_1 and P_2 vertically as shown and view from the other side AC. The images P_1' and P_2' of pins P_1 and P_2 will be seen. Fix two more pins Q_1 and Q_2 such that they are in line with the images P_1' and P_2'. Remove the prism. Draw a straight line PQ joining P_1 and P_2. This is the incident ray. Draw a normal to the face AB at Q. The angle i is the angle of incidence. Now draw a straight line RS joining pins Q_1 and Q_2. RS represents the emergent ray. Produce PQ and SR. They meet at O. The angle of deviation δ can now be measured.

Repeat the experiment for various inclinations of the line PQ, i.e. for various angles of incidence i and measure the angle δ in each case. It is found that as the angle of incidence increases, the angle of deviation decreases at first, becoming minimum for a certain value i_0 of the angle of incidence. Beyond this angle of incidence, the angle of deviation increases with increase in the angle of incidence [see Fig. 11.25(b)].

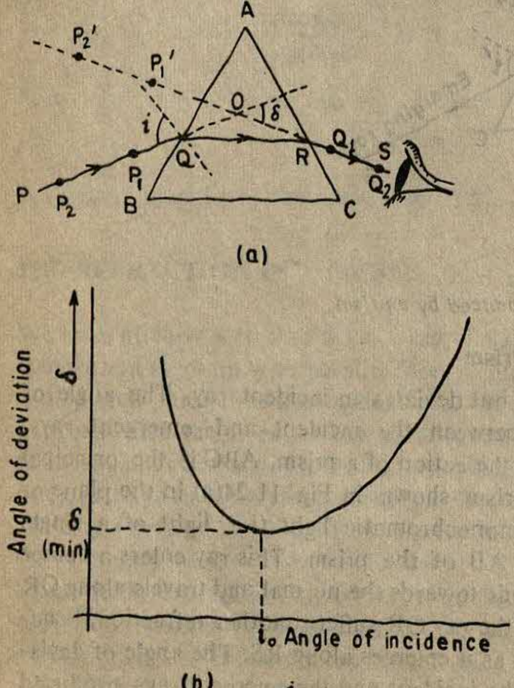


Fig. 11.25 *(a) Tracing the course of rays through a prism*
(b) Variation of angle of deviation with angle of incidence

Measurement of Refractive Index of the Prism Glass

The fact that the angle of deviation becomes minimum at a certain angle of incidence provides an accurate method for measuring the refractive index of the material of the prism.

Let us again refer to Fig. 11.24(b). From quadrilateral AQNR

$$A + \alpha = 180°$$

From triangle QNR,

$$r + r' + \alpha = 180°$$

Therefore $$A = r + r'$$

Now, from triangle MQR, we have

$$\begin{aligned}\delta &= (i - r) + (i' - r') \\ &= (i + i') - (r + r') \\ &= (i + i') - A \qquad (11.5)\end{aligned}$$

Now, when the deviation δ is minimum $(= \delta_m)$, the experiment of Fig. 11.25(a) shows that $i = i'$, i.e. the incident and the emergent rays are symmetrical with respect to the refracting faces AB and AC and the refracted ray QR in the prism is parallel to the base BC. When $i = i'$, $r = r'$ so that $A = 2r$ or $r = A/2$. From Eq. (11.5) we have, at minimum deviation

$$\delta_m = 2i - A$$

or $$i = \frac{1}{2}(A + \delta_m)$$

From Snell's law, the refractive index of the material of the prism is given by

$$n = \frac{\sin i}{\sin r} = \frac{\sin\left[\frac{1}{2}(A + \delta_m)\right]}{\sin\left(\frac{A}{2}\right)} \qquad (11.6)$$

Knowing the values of A and δ_m, the value of n is determined.

EXAMPLE 11.10 A prism produces a minimum deviation of 51° for a certain angle of incidence. The same prism produces a deviation of 62° 48′ for two values of the angles of incidence, namely, 40° 6′ and 82° 42′. Find (a) the refracting angle of the prism, (b) the angle of incidence at minimum deviation, and (c) the refractive index of the prism.

Solution: (a) The incident ray is deviated through $\delta = 62° \, 48'$ when angle $i = 40° \, 6'$. It is clear from Fig. 11.24(b) that if the emergent ray is reversed, i.e. if the ray travels along SR, it will emerge from face AB along QP, the deviation δ remaining the same. Hence $i' = 82° \, 42'$. Now

$$\delta = i + i' - A$$

or

$$\begin{aligned} A &= i + i' - \delta \\ &= 40° \, 6' + 82° \, 42' - 62° \, 48' \\ &= 60° \end{aligned}$$

(b) For minimum deviation, $i = i'$. Hence,

$$\delta_m = 2i - A$$

or

$$\begin{aligned} i &= \frac{1}{2}(A + \delta_m) \\ &= \frac{1}{2}(60° + 51°) \\ &= 55° \, 30' \end{aligned}$$

(c)

$$\begin{aligned} n &= \frac{\sin\left[\frac{1}{2}(A + \delta_m)\right]}{\sin(A/2)} \\ &= \frac{\sin\left[\frac{1}{2}(60° + 51°)\right]}{\sin(60°/2)} \\ &= 1.65 \end{aligned}$$

11.8 REFRACTION AT A SPHERICAL SURFACE

Let us now consider refraction at a spherical surface. At each point on the surface, the refracted ray obeys the Snell's law of refraction. In Fig. 11.26 LMN represents a convex refracting surface of radius of curvature R of a medium of refractive index n_2. Consider a point object O on the principal axis situated in an optically less dense medium of refractive index n_1 ($n_1 < n_2$). A ray OA incident at an angle i is refracted along AI subtending an angle r with the normal AC at point A; C is the centre of curvature of the surface. Angles i and r are

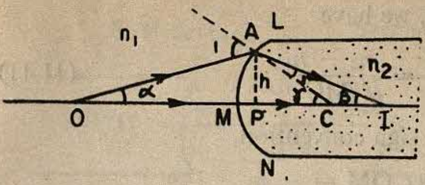


Fig. 11.26 *Refraction at a convex refracting surface ($n_2 > n_1$)*

related by Snell's law:

$$\frac{n_2}{n_1} = \frac{\sin i}{\sin r} \tag{11.7}$$

Another ray OM falling normally on the surface goes through undeviated. The two refracted rays meet at point I which is the image of object O.

We now obtain the relation between the object and image distances. In triangle AOC, i is the exterior angle so that

$$i = \alpha + \gamma \tag{11.8}$$

Similarly for triangle AIC

$$\gamma = r + \beta \tag{11.9}$$

We now make the simplifying assumption that the rays are paraxial, i.e. A is close to M, so that angles i, r, α, β, and γ are small. For small angles (expressed in radians) $\sin i \simeq i$ and $\sin r \simeq r$, so that Eq. (11.7) implifies to

$$\frac{n_2}{n_1} = \frac{i}{r}$$

or

$$n_1 i = n_2 r$$

Using Eqs (11.8) and (11.9), we have

$$n_1(\alpha + \gamma) = n_2(\gamma - \beta)$$

or

$$n_2\beta + n_1\alpha = (n_2 - n_1)\gamma \tag{11.10}$$

From A draw AP perpendicular to the principal axis and let AP $= h$. Now in triangles AOP, AIP and ACP we have

$$\tan \alpha = \frac{h}{\text{OP}} \qquad \tan \beta = \frac{h}{\text{PI}} \qquad \tan \gamma = \frac{h}{\text{PC}} \tag{11.11}$$

if A is close to M, OP $\simeq$ OM, PI $\simeq$ MI and PC $\simeq$ MC. Also, angles α, β and γ will be small so that $\tan \alpha \simeq \alpha$ ($\because \sin \alpha \simeq \alpha$, $\cos \alpha \simeq 1$ for small α), $\tan \beta \simeq \beta$ and $\tan \gamma \simeq \gamma$. Therefore Eqs (11.11) become

$$\alpha = \frac{h}{\text{OM}} \qquad \beta = \frac{h}{\text{MI}} \qquad \gamma = \frac{h}{\text{MC}}$$

Using these values in Eq. (11.10), we have

$$\frac{n_2}{\text{MI}} + \frac{n_1}{\text{OM}} = \frac{n_2 - n_1}{\text{MC}} \tag{11.11}$$

Now, according to the Cartesian sign convention,

Object distancc OM $= -u$

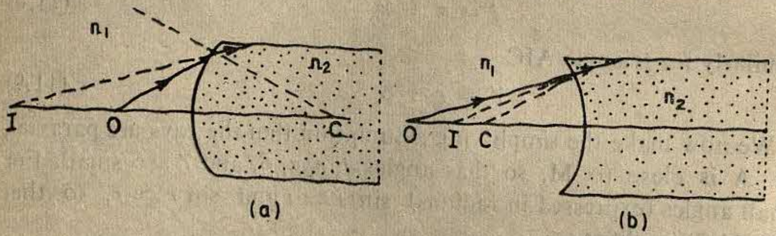
Image distance MI $= +v$

and Radius of curvature MC $= +R$

Therefore, with the signs included, Eq. (11.11) becomes

$$\frac{n_2}{v} - \frac{n_1}{u} = \frac{n_2 - n_1}{R} \tag{11.12}$$

We have derived Eq. (11.12) for a convex refracting surface forming a real image (Fig. 11.26). This relation also holds for a convex surface forming a virtual image [Fig. 11.27(a)] and a concave refracting surface [Fig. 11.27(b)]. This has been left as an exercise for the students.



(a) (b)

Fig. 11.27 *(a) Convex refracting surface*
(b) Concave refracting surface

Note: If the object is placed in medium of refractive index n_2 so that the incident rays are in medium n_2 and refracted rays are in medium n_1, then in Eq. (11.12), n_1 and n_2 are interchanged and we have

$$\frac{n_1}{v} - \frac{n_2}{u} = \frac{n_1 - n_2}{R} \tag{11.13}$$

EXAMPLE 11.11 Two concave glass refracting surfaces, each with radius of curvature $R = 35$ cm and refractive index $n = 1.5$ are placed facing each other in air as shown in Fig. 11.28. A point object O is placed at a distance of $R/2$ from one of the surfaces as shown. Find the separation between the images I_1 and I_2 of O formed by each refracting surface.

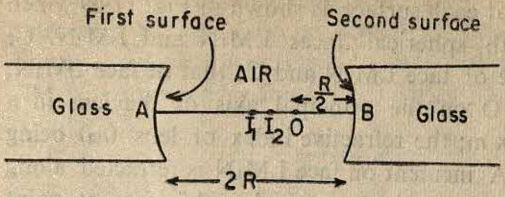


Fig. 11.28

Solution: For refraction at the first surface (incident rays from O are from right to left)

$$u = OA = -\frac{3}{2}R = -\frac{105}{2}\text{ cm}$$

$$n_2 = 1.5 \qquad n_1 = 1 \qquad R = -35\text{ cm}$$

The distance v_1 of image I_1 from A is given by the relation

$$\frac{n_2}{v_1} - \frac{n_1}{u} = \frac{n_2 - n_1}{R}$$

or

$$\frac{1.5}{v_1} - \frac{1}{-105/2} = \frac{1.5 - 1}{-35}$$

which gives

$$v_1 = -45\text{ cm} \quad \text{or} \quad AI_1 = 45\text{ cm}$$

For refraction at the second surface (incident rays from O are from left to right)

$$u = OB = -\frac{R}{2} = -\frac{35}{2}\text{ cm}$$

The distance v_2 of image I_2 from B is given by

$$\frac{n_2}{v_2} - \frac{n_1}{u} = \frac{n_2 - n_1}{R}$$

or

$$\frac{1.5}{v_2} - \frac{1}{-35/2} = \frac{1.5 - 1}{-35}$$

which gives $v_2 = -21$ cm or $BI_2 = 21$ cm. Therefore $AI_2 = AB - BI_2 = 70\text{–}21 = 49$ cm. Hence the separation between I_1 and I_2 is

$$I_1I_2 = AI_2 - AI_1 = 49\text{–}45 = 4.0\text{ cm}$$

11.9 REFRACTION IN A THIN LENS

A lens is a combination of two refracting surfaces. The image is formed after two refractions, one at each surface as shown in Fig. 11.29 which shows a convex lens with spherical faces LM$_1$N and LM$_2$N. C$_1$ is the centre of curvature of face LM$_1$N and C$_2$ that of face LM$_2$N. Consider a point object O on the principal axis of the lens in a medium of refractive index n_1; the refractive index of lens (n_2) being greater than n_1. A ray OA incident on face LM$_1$N is refracted along AB by this surface, bending towards the normal to this face at point A. If the second face LM$_2$N were absent, the ray AB would have intersected the principal axis at I′ so that I′ in the image of O due to refraction at face LM$_1$N. But due to the second face LM$_2$N, the ray AB incident on it will be refracted along BC, bending away from the normal to the face LM$_2$N at point B. The emergent ray BC intersects the principal axis at point I which is the image of O formed by the lens.

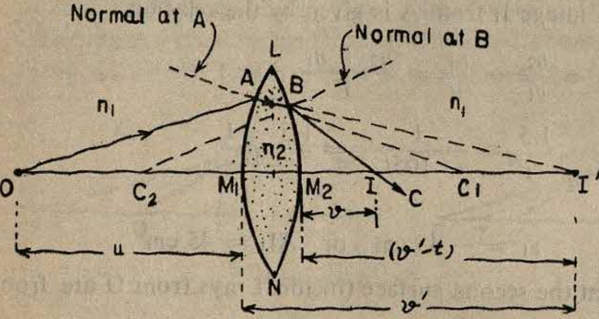


Fig. 11.29 *Refraction through a convex lens*

The Thin Lens Formula

Consider a lens of refractive index n_2 in a medium of refractive index n_1 as shown in Fig. 11.29. We assume that the lens is thin so that its thickness M$_1$M$_2$ can be neglected in comparison to the object and image distances and the radii of curvature of the two faces of the lens. We also assume that the rays are paraxial, i.e. they subtend very small angles with the normals to the faces and with the principal axis; this assumption is satisfied if the aperture LN of the lens is small. If these two assumptions are satisfied, a simple relationship is obtained between the object distance (u), the image distance (v) and the focal length (f) of the lens. We obtain this relation by considering refractions at the two faces of the lens.

Refraction at Face LM$_1$N

For refraction at this face, the incident OA is in medium n_1 and the refracted ray AB is in medium n_2 and I′ is the image of O due to refraction at this face, as explained above. Therefore, the object distance (u), image distance (v') and radius of curvature (R_1) of face LM_1N are related by Eq. (11.12) which now reads

$$\frac{n_2}{v'} - \frac{n_1}{u} = \frac{n_2 - n_1}{R_1} \tag{11.14}$$

Refraction at Face LM_2N

For refraction at this face, the incident ray AB is in medium n_2 and the refracted ray BC is in medium n_1. For this face I′ serves as the virtual object and I is the image of I′ due to refraction at this face. Therefore, for this face, the object distance $= I'M_2 = I'M_1 - M_1M_2 = v' - t$, where $t = M_1M_2$ is the thickness of the lens and the image distance $= IM_2 = v$. Let R_2 be the radius of curvature of the face LM_2N. The object distance ($v' - t$), the image distance (v) and the radius of curvature (R_2) are related by Eq. (11.13) which now reads

$$\frac{n_1}{v} - \frac{n_2}{v' - t} = \frac{n_1 - n_2}{R_2}$$

If the lens is thin, i.e. $t \ll v'$, then $(v' - t) \simeq v'$ and this relation becomes

$$\frac{n_1}{v} - \frac{n_2}{v'} = \frac{n_1 - n_2}{R_2} \tag{11.15}$$

Remember that the relations (11.14) and (11.15) already have the sign convention incorporated in them. Adding Eqs. (11.14) and (11.15), we get

$$\frac{n_1}{v} - \frac{n_1}{u} = (n_2 - n_1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right)$$

or
$$\frac{1}{v} - \frac{1}{u} = \left(\frac{n_2 - n_1}{n_1}\right)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \tag{11.16}$$

Focal Length of a Thin Lens

If the object is at infinity ($u = \infty$), the image is formed at the focus, i.e. $v = f$. Putting $u = \infty$ and $v = f$ in Eq. (11.16) gives

$$\frac{1}{f} = \left(\frac{n_2 - n_1}{n_1}\right)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \tag{11.17}$$

Thus we find that the focal length (f) of a lens depends upon its refractive index (n_2), the refractive index (n_1) of the medium in which it is placed and the radii of curvature (R_1 and R_2) of its curved faces. Relation (11.17) is called the *lens maker's formula.*

Combining Eqs (11.16) and (11.17), we have

$$\frac{1}{v}-\frac{1}{u}=\frac{1}{f}=\left(\frac{n_2-n_1}{n_1}\right)\left(\frac{1}{R_1}-\frac{1}{R_2}\right) \qquad (11.18)$$

Note: The relation (11.18) holds for paraxial rays and thin lenses. This relation includes the cartesian sign conventions regarding the distances u, v, f, R_1 and R_2. According this sign convention, u is negative, v is positive for real images, and negative for virtual images, f of a convex lens is positive and is negative for a concave lens, $R_1 = M_1C_1$ is positive and $R_2 = M_2C_2$ is negative for a convex lens. For a concave lens R_1 is negative and R_2 is positive. Equation (11.18) has been derived for a convex lens forming a real image. The same relation is applicable to a convex lens forming a virtual image and for a concave lens, which forms only virtual images. This has been left as an exercise for the students.

Linear Magnification

Figure 11.30 shows the ray diagram of a convex lens forming a real image A′B′ of an extended (small) object AB.

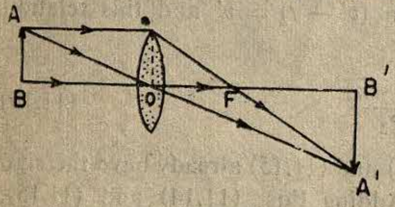


Fig. 11.30

From similar triangles AOB and A′OB′ we have

$$\frac{A'B'}{AB}=\frac{OB'}{OB}$$

From our sign convention, $A'B' = -h_2$, $AB = +h_1$, $OB' = +v$ and $OB = -u$, substituting these values in the above equation, we have

$$\frac{-h_2}{+h_1}=\frac{+v}{-u}$$

or

$$\frac{h_2}{h_1}=+\frac{v}{u}$$

Thus,

$$\text{Linear magnification } m(=h_2/h_1)=\frac{v}{u} \qquad (11.19)$$

This relation holds for both convex and concave lenses and for real as well as virtual images.

Power of a Lens

The power (P) of a lens is a characteristic of the focal properties of a lens and is defined as the reciprocal of the focal length (expressed in metres), i.e.

$$P = \frac{1}{f\,(\text{in metres})} \tag{11.20}$$

The unit of power is m^{-1} which is called the *dioptre* (symbol D). Thus a convex lens of focal length 50 cm ($f = +0.5$ m) has a power of $1/+0.5$ m $= +2.0$ D. The power of a concave lens of focal length 200 cm ($f = -2.0$ m) is $1/-2.0$ m $= -0.5$ D.

EXAMPLE 11.12 A convex lens of crown glass ($n_g = 1.5$) has a focal length of 15 cm. The lens is placed in: (a) water ($n_w = 1.33$) and (b) carbon bisulphide ($n_c = 1.65$). Determine in each case, whether the lens behaves as a converging or diverging lens and determine its focal length.

Solution: For lens in air

$$\frac{1}{15} = (n_g - 1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) = 0.5\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \tag{i}$$

For lens in water,

$$\frac{1}{f_w} = \frac{(n_g - n_w)}{n_w}\left(\frac{1}{R_1} - \frac{1}{R_2}\right) = \frac{0.17}{1.33}\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \tag{ii}$$

For lens in carbon bisulphide,

$$\frac{1}{f_c} = \frac{(n_g - n_c)}{n.}\left(\frac{1}{R_1} - \frac{1}{R_2}\right) = -\frac{0.15}{1.65}\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \tag{iii}$$

Dividing (i) by (ii), we get

$$f_w = +58.7 \text{ cm}$$

The positive sign indicates that the lens is converging. Dividing (i) by (iii), we get

$$f_c = -82.5 \text{ cm}$$

The negative sign indicates that the lens behaves as a diverging lens when it is immersed in carbon bisulphide.

EXAMPLE 11.13 A luminous object and a screen are mounted on an optical bench and a converging lens is placed between them so that a sharp image is received on the screen; the linear magnification of the image is found to be 2.5. The lens is now moved 30 cm nearer the screen and a sharp image is again formed. Find the focal length of the lens.

Solution: If O and I are the object and the screen respectively and L_1 and L_2 are the two conjugate positions of the lens (Fig. 11.31), then

$$OL_1 = IL_2$$

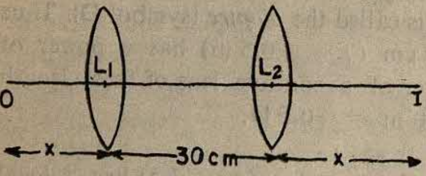


Fig. 11.31

For lens at L_1, we have $u = -x$ and $v = 30 + x$

Linear magnification $m = \frac{v}{u} = -2.5$ (because the image is inverted, being real)

Thus $$\frac{30 + x}{-x} = -2.5$$

or $$x = 20 \text{ cm}$$

Hence $u = -20$ cm and $v = 30 + 20 = 50$ cm

Substituting in $\frac{1}{v} - \frac{1}{u} = \frac{1}{f}$, we have

$$\frac{1}{50} - \frac{1}{-20} = \frac{1}{f}$$

which gives $f = 14.3$ cm.

EXAMPLE 11.14 A convex lens forms a real image of an object on a screen; the magnification of the image being 3/2. The object and the screen are kept fixed and the lens is moved through a distance of 16 cm when a sharp image is again formed on the screen; the magnification now being 2/3. Find the focal length of the lens.

Solution: Refer again to Fig. 11.31. The separation L_1L_2 now is 16 cm. For lens at L_1, we have

$$u = -x \qquad v = 16 + x \qquad m_1 = -\frac{3}{2}$$

Now $$m_1 = \frac{v}{u} = -\frac{3}{2}$$

or $$\frac{16 + x}{-x} = -\frac{3}{2}$$

or $$x = 32 \text{ cm}$$

Therefore $u = -32$ cm and $v = +48$ cm. Substituting these values in the lens formula we get

$$f = 19.2 \text{ cm}$$

11.10 COMBINATIONS OF THIN LENSES

Two or more lenses are often combined in various optical instruments. The purpose is: (i) to increase the magnification of the image, (ii) to make the final image erect and (iii) to minimize certain defects (aberrations) of the image. The location, size and nature of the final image produced by a combination of lenses can be found by two methods: (i) by using the lens formula and (ii) by drawing the ray diagram. In both the methods, we first find the image formed by the first lens, treat this image as the object for the second lens and find the image due to the second lens, this image then serves as the object for the third lens and so on. This procedure is illustrated in Examples 11.15 and 11.16 below.

When lenses are used in combination, each lens magnifies the image formed by the preceding lens. Hence the total magnification (m) of the final image is given by the product of the magnifications due to individual lenses, i.e.

$$m = m_1 \times m_2 \times m_3 \times \ldots \tag{11.21}$$

EXAMPLE 11.15 An object O is placed 15 cm from a convex lens A of focal length 10 cm and its image, I_1 is formed on a screen on the other side of the lens. A concave lens B is now placed midway between A and I_1 and the screen is moved back 10 cm to receive a clear image I_2. Find: (a) the focal length of lens B, and (b) the linear magnification of the final image I_2.

Solution: (a) For lens A, $u = \text{PO} = -15$ cm, $f = +10$ cm and $v = \text{PI}_1$? (see Fig. 11.32).

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f}$$

$$\frac{1}{v} = \frac{1}{f} + \frac{1}{u} = \frac{1}{10} - \frac{1}{15}$$

which gives $$v = +30 \text{ cm} = \text{PI}_1$$

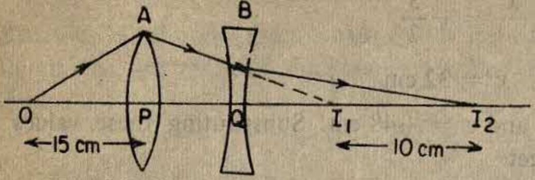


Fig. 11.32

The image I_1 serves as the virtual object for lens B which is placed midway between P and I_1. Thus for lens B,

$$u = QI_1 = \frac{1}{2}(PI_1) = +15 \text{ cm}$$

$$v = QI_2 = QI_1 + I_1I_2 = 15 + 10 = 25 \text{ cm}$$

The focal length f of lens B is obtained from the formula

$$\frac{1}{f} = \frac{1}{v} - \frac{1}{u}$$

$$= \frac{1}{25} - \frac{1}{15}$$

which gives $f = -37.5$ cm

The negative sign is expected since lens B is concave.

(b) Linear magnification of $I_2 = m_A \times m_B$, where m_A and m_B are magnifications due to A and B respectively.

For lens A,

$$m_A = \frac{v}{u} = \frac{30}{-15} = -2$$

For lens B,

$$m_B = \frac{v}{u} = \frac{25}{15} = \frac{5}{3}$$

$$\text{Total magnification} = -2 \times 5/3 = -10/3$$

The negative sign here indicates that the final image is inverted.

EXAMPLE 11.16 Two convex lenses A and B of focal lengths 20 cm and 10 cm are placed coaxially 10 cm apart. An object is placed on the common axis at a distance of 10 cm from lens A. Find the position and magnification of the final image.

Solution: Referring to Fig. 11.33 we have, for lens A,

$$f_1 = +20 \text{ cm} \qquad u_1 = -10 \text{ cm}$$

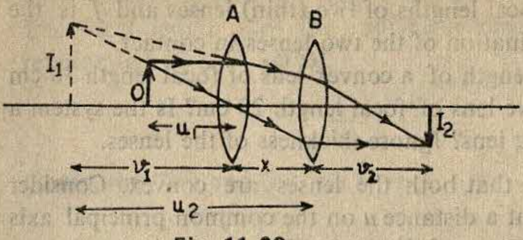


Fig. 11.33

The image distance v_1 is given by

$$\frac{1}{v_1} = \frac{1}{f_1} + \frac{1}{u_1}$$

which gives $v_1 = -20$ cm.

Thus a virtual image is formed at I_1 at a distance of 20 cm from lens A, if the lens B were absent. This image acts as a virtual object for lens B which forms the final image at I_2 at a distance v_2 from lens B.

For lens B we have, since $x = 10$ cm

$$u_2 = -(20 + 10) = -30 \text{ cm}$$

$$f_2 = +10 \text{ cm}$$

The image distance v_2 is given by

$$\frac{1}{v_2} = \frac{1}{f_2} + \frac{1}{u_2} = \frac{1}{10} - \frac{1}{30} = \frac{1}{15}$$

which gives $v_2 = +15$ cm

Thus a real image I_2 is formed at a distance of 15 cm from lens B.

$$\text{Magnification due to A}(m_1) = \frac{v_1}{u_1} = \frac{-20}{-10} = +2$$

$$\text{Magnification due to B}(m_2) = \frac{v_2}{u_2} = \frac{+15}{-30} = -\frac{1}{2}$$

Magnification of the final image is

$$m = m_1 \times m_2 = 2 \times \left(-\frac{1}{2}\right) = -1$$

This shows that the final image is inverted and is of the same size as the object.

EXAMPLE 11.17 (a) Derive the relation

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2}$$

where f_1 and f_2 are the focal lengths of two (thin) lenses and f is the focal length of the combination of the two lenses in contact.

(b) What is the focal length of a convex lens of focal length 30 cm in contact with a concave lens of focal length 20 cm? Is the system a converging or a diverging lens? Ignore thickness of the lenses.

Solution: Let us assume that both the lenses are convex. Consider a point object A placed at a distance u on the common principal axis [Fig. 11.34(a)]. In the absence of the second lens, let its image due to the first lens be formed at C. Let its distance from the first lens be v'. From the lens formula, we have

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f_1} \tag{i}$$

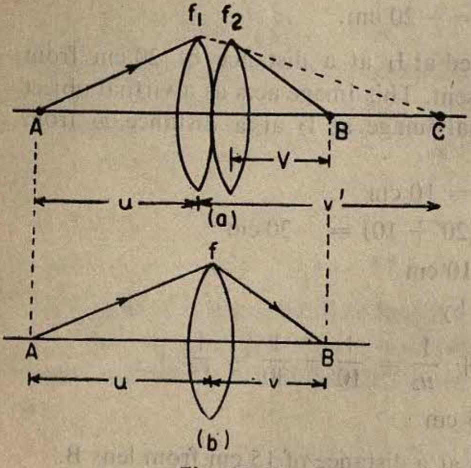


Fig. 11.34

If the second lens is also present, the final image is formed at B. For this lens, C may be regarded as the virtual object and B its image. Thus, for the second lens, the object distance is v' and image distance is v, so that

$$\frac{1}{v} - \frac{1}{v'} = \frac{1}{f_2} \tag{ii}$$

Adding (i) and (ii), we get

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f_1} + \frac{1}{f_2} \tag{iii}$$

Now suppose we replace the combination of lenses by a single lens of a suitable focal length f such that it forms the image at B of an object at A [Fig. 11.34(b)]. This lens would then be the equivalent lens because it has the same effect as the two lenses together have. The focal length f of the equivalent lens must obviously satisfy the relation [see Fig. 11.34(b)].

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f} \tag{iv}$$

Comparing (iii) and (iv), we have

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2}$$

(b) $f_1 = +30$ cm (convex lens)

$f_2 = -20$ cm (concave lens)

The focal length of the combination is determined as follows.

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2}$$

$$= \frac{1}{30} - \frac{1}{20}$$

$$= -\frac{1}{60}$$

or $$f = -60 \text{ cm}$$

The negative sign shows that the combination is a concave or diverging lens of focal length 60 cm.

11.11 DISPERSION OF LIGHT

We have learnt that the refraction of light occurs because the velocity of light changes as it travels from one medium into another. The speed of light in any material medium is less than that in a vacuum. The ratio $n = c/v$ is called the index of refraction (or refractive index) of the medium where c and v are respectively the speeds of light in vacuum and in the medium. When white light enters a prism, it is refracted, but the constituent colours of white light are refracted unequally as shown in Fig. 11.35. The red is refracted (or deviated) the least and the violet is refracted the most. From Snell's law $n = \sin i/\sin r$, the refractive index for red is less than that for violet. In other words, the speed of red light is higher than that of violet light. Thus, a medium does not have one definite index of refraction; it has a range of indices corresponding to a range of colours of light. Since each colour has its own characteristic wavelength (or frequency), the refractive index of a medium will be different for different wavelengths. The

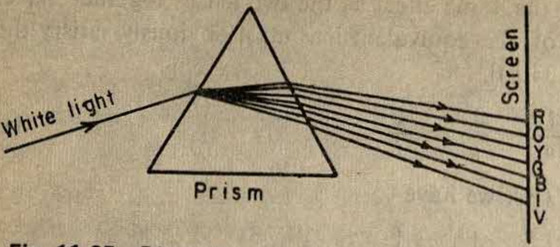


Fig. 11.35 *Dispersion of white light by a prism: Red, Orange, Yellow, Green, Blue, Indigo, Violet (ROYGBIV)*

variation of the refractive index of a medium (and hence of the velocity of light in the medium) with the wavelength (or frequency) is referred to as *dispersion*. The prism disperses the colours of white light and we say that we have got the "spectrum" of the light.

The index of refraction of a material can be determined for monochromatic light by taking a prism of that material and measuring its angle A and the minimum deviation angle δ_m. This is accomplished with the help of an optical instrument called the spectrometer.

The Spectrometer (or Spectroscope)

In the experiment shown in Fig. 11.34, the different colours overlap and we get what is called an impure spectrum. To produce a pure spectrum, in which the colours are distinctly separated, we use a spectrometer. The essential parts of a spectrometer, illustrated in Fig. 11.36 are: (i) a collimator, (ii) a prism table, and (iii) a telescope.

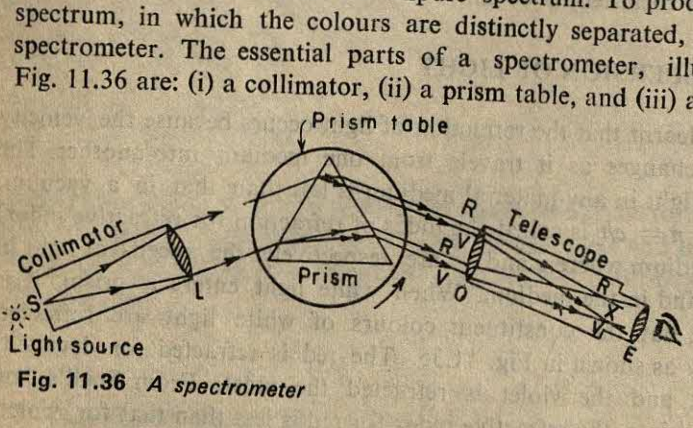


Fig. 11.36 *A spectrometer*

The collimator is a tube having a convex lens L at one end and an adjustable vertical slit S at the other. The slit can slide in the tube so that it can be placed at the focus of the lens. Remember the focal length of a lens is different for different colours. The prism table is a circular horizontal base for the prism, which can be rotated about a vertical axis which passes through its centre. The horizontal axis of the collimator tube and the vertical axis of the prism table intersect each other. The telescope is capable of rotation about a central vertical axis with its horizontal axis passing through the vertical axis of the prism table. The position of the telescope is described by means of a graduated circular scale whose centre is the axis of rotation. The lenses O and E are the objective and the eyepiece of the telescope. The cross-wire is fixed in front of the eyepiece at the position marked X where the spectrum is produced.

The slit is illuminated with the light from a source. The slit is made as narrow as possible and the distance SL is adjusted till S is at the focal plane of the lens L. A parallel beam of light emerging from the lens L is made to fall on the prism. The prism produces dispersion. The rays of a given colour emerging from the prism are parallel. They are received by a telescope which is already focussed at infinity. The coloured rays converge at the focal plane of the objective. A cross-wire is fixed at this position marked X where the spectrum is produced. The spectrum is viewed through the eyepiece E. The spectrum consists of the coloured images of slit S. If the slit is narrow, these images do not overlap and a pure spectrum is obtained. The prism is placed in the minimum-deviation position and the vertical cross-wire is made to coincide with one of the coloured images. The position of the telescope is measured on the circular scale. The telescope is then rotated so as to receive direct light and its position is again noted. The difference between the two positions of the telescope gives the minimum deviation angle δ_m for that colour. Knowing the angle A of the prism, its refractive index for that colour is determined by using the relation (11.6).

Table 11.2 gives us an idea as to the variation of refractive index with the wavelength of light.

Table 11.2 Indices of refraction for crown glass and flint glass at Five Typical Wavelengths

Wavelength Å	*Crown glass*	*Flint glass*
6563 (red)	1.515	1.644
5890 (yellow)	1.517	1.650
5461 (green)	1.519	1.655
4861 (blue)	1.523	1.664
4047 (violet)	1.532	1.685

Direct Vision Spectroscope

Figure 11.37 shows a simple form of the direct-vision spectroscope in which three crown glass and two flint glass prisms of suitable angles are mounted alternately in a brass tube and held in position by cork pieces. They are arranged with their refracting angles in opposite directions so that the middle part of the spectrum suffers no deviation. The prisms are cemented together by canada balsam to avoid loss of intensity due to reflection at the faces. There is an adjustable slit S placed parallel to the refracting edges of the component prisms and an achromatic lens at one end of the tube and an eyepiece at the other.

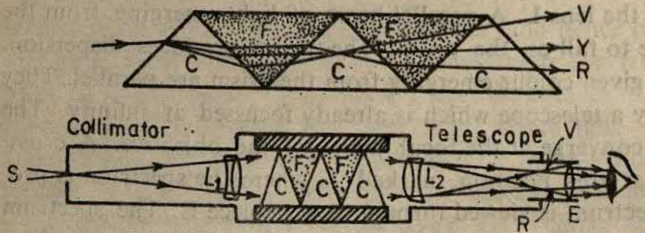


Fig. 11.37 *Direct vision spectroscope*

A narrow beam of white light passing through the slit is rendered parallel by the lens L_1. It then traverses the combination of the prisms, where it is dispersed. The paths of the red and blue rays are shown in the figure and it may be noted that the rays cross one another in the second crown glass prism. They emerge out as a system of parallel beam which falling on another converging lens L_2 is brought to focus in its second focal plane to form a spectrum VR. This spectrum is next examined by the eyepiece E which acts as a simple magnifying glass.

The direct vision spectroscope is a pocket-sized instrument. By directing it towards a star or a flame, a quick qualitative study of the spectra can be made.

Dispersive Power

The ability of a prism to deviate the different colours of a composite light along different directions is characterized by the dispersive power of the material of which the prism is made. It is denoted by ω and is defined as

$$\omega = \frac{n_b - n_r}{n_y - 1} \tag{11.22}$$

where n_b, n_r and n_y are the refractive indices of the material of the prism for blue, red and yellow lights respectively.

EXAMPLE 11.18 Using a spectrometer, the following data are obtained for crown glass and flint glass prisms:

Crown glass prism

Angle of the prism,

$$A = 72.0°$$

Minimum deviation angle

$$\delta_b = 54.6°$$
$$\delta_r = 53.0°$$
$$\delta_y = 54.0°$$

Flint glass prism

$$A = 60.0°$$

$$\delta_b = 52.8° \qquad \delta_r = 50.6° \qquad \delta_y = 51.9°$$

b, *r* and *y* refer to particular wavelengths in the blue, red and yellow bands. Compare the dispersive powers of the two varieties of glass prisms.

Solution:

$$n = \frac{\sin\left[\frac{1}{2}(A + \delta_m)\right]}{\sin\left(\frac{A}{2}\right)} \tag{i}$$

Crown glass Substitute the values of A and δ_m (for each colour) (i) and calculate n for each colour. The results are

$$n_b = 1.520 \qquad n_r = 1.509 \qquad n_y = 1.516$$

The dispersive power of crown glass is

$$\omega_c = \frac{n_b - n_r}{n_y - 1} = \frac{1.520 - 1.509}{1.516 - 1} = 0.0213$$

Flint glass The data for flint glass gives:

$$n_b = 1.666 \qquad n_r = 1.644 \qquad n_y = 1.657$$

which give

$$\omega_f = 0.0355$$

Therefore

$$\frac{\omega_f}{\omega_c} = \frac{0.0355}{0.0213} = 1.57$$

Since flint glass has a higher dispersive power than crown glass, a flint glass prism gives a better spread out spectrum than a crown glass prism.

11.12 TYPES OF SPECTRA

Spectra may be divided into two principal classes known as emission spectra and absorption spectra. When light emitted by a source is examined with a spectroscope, we obtain an emission spectrum. Absorption spectrum is obtained when the light emitted from a source is made to pass through an absorbing material and then examined with a spectroscope. Every source has its characteristic emission spectrum and every substance has its characteristic absorption spectrum. The analysis of these spectra, therefore, gives a very valuable insight into the structure of the substances.

Emission Spectra

There are three types of emission spectra, namely, continuous, line and band spectra.

Continuous Spectrum A continuous spectrum consists of a wide range of unseparated wavelengths. Solids at high temperature (and fairly dense luminous gases under very special conditions) give continuous spectra. Such spectra are not characteristic of the emitting substance, but depend only on the temperature. In Fig. 11.38 three such spectra at temperatures of 1000, 2000 and 4000°C are shown.

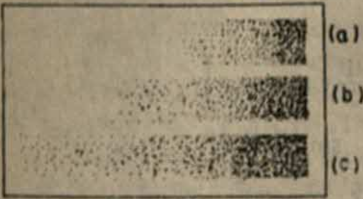


Fig. 11.38 *Continuous spectra (negatives) of (a) a solid at 1000 °C; (b) at 2000 °C; and (c at 4000 °C*

Line Spectrum Unlike continuous spectra which are emitted by the matter in bulk, the line spectra are emitted by the atoms in the gaseous state. The substances in the atomic state, such as in the sodium vapour lamp, mercury vapour lamp and in gases in vacuum discharge tubes, give line spectra. These are sharp lines of definite wavelength, with no light emitted at the intermediate wavelengths. Since free atoms emit their own characteristic light, the line spectrum, unlike the continuous spectrum, is a characteristic of the emitting substance. Therefore, the line spectrum of a gas can be used as a means of identifying the gas. The line spectra of mercury, helium, and neon are shown in Fig. 11.39.

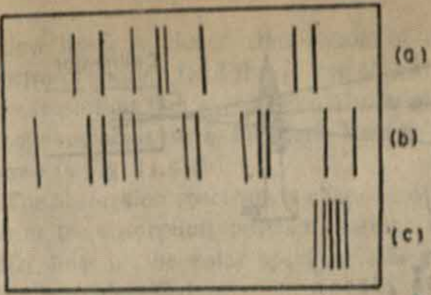


Fig. 11.39 *Line spectra of: (a) mercury in a glass enclosure, (b) helium in a glass discharge tube, and (c) neon in a glass discharge tube*

Band Spectrum Band spectra consist of a number of bright bands, each band having one sharp edge and the other edge fading gradually. On closer examination, using a good spectroscope, each band is found to consist of several fine lines which are very close to each other at the sharp edge of the band but are well separated near the diffuse end. Remember that the continuous spectra will not show these separate lines even with the best spectroscopes. The band spectra arise from molecules which are combinations of two or more atoms. Calcium or barium salts in a Bunsen burner flame or in an arc, and gases like carbon dioxide and nitrogen in the molecular state in a vacuum tube, give band spectra. Figure 11.40 shows the band spectrum of the air in a discharge tube at low pressure.

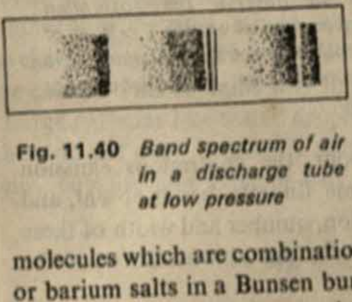


Fig. 11.40 *Band spectrum of air in a discharge tube at low pressure*

Absorption Spectra

If a substance is placed between a source emitting a continuous range of wavelengths and the slit of the spectroscope, we get dark regions in the spectrum. These dark regions are due to the absorption of certain wavelengths by the substance. We say that we have obtained the absorption spectrum of the substance. As in the case of emission spectra, we may classify absorption spectra as: (i) continuous, (ii) line, and (iii) band absorption spectra.

A pure green glass, which absorbs everything except green, gives a continuous absorption spectrum when placed in the path of white light. If white light is passed through dilute solutions of blood or

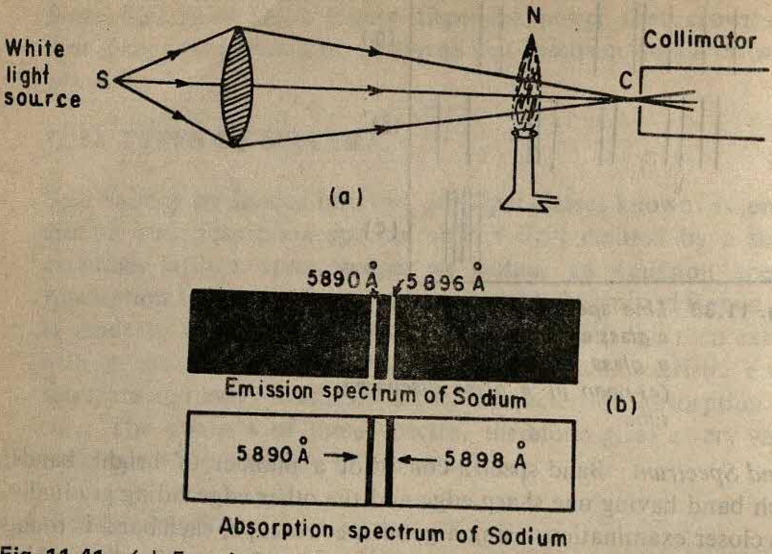


Fig. 11.41 *(a) Experimental arrangement to observe the absorption spectrum of a substance, temperature of source S > that of flame (b) Emission and absorption spectra of sodium (The background continuous spectrum is, for convenience, not shown in the absorption spectrum)*

through a molecular gas like iodine vapour, the continuous emission spectrum of white light will have some discrete bands absent, and these regions are dark bands. The position, number and width of these bands, in the band absorption spectrum, are characteristic of the absorbing substance.

- If white light is passed through an absorbing gas or vapour, the absorption spectrum consists of a number of dark lines across a continuous background. The solar spectrum is an excellent example of a line absorption spectrum. The dark lines in the solar spectrum are called *Fraunhofer lines*. Certain wave-lengths of the continuous spectrum of light from the central core of the sun are absorbed by the sun's atmosphere which contains gases and many elements, like sodium and iron, in the gaseous form.

Figure 11.41(a) illustrates a simple experiment to obtain the line absorption spectrum of a substance (say sodium) in the vapour form. White light from a source S is passed through a sodium flame N and examined with a spectroscope, C being its collimator. The absorption spectrum shows a pair of dark lines across a continuous emission spectrum of S. When the source S is shut off we see the emission spectrum of sodium vapours. This spectrum consists of two bright yellow lines. A closer examination of the absorption and emission spectra of sodium (and this is true of all other substances) reveals a very important fact about light. The bright emission lines of sodium exactly coincide with the dark lines in its absorption spectrum as shown in Fig. 11.41(b).

The absorption spectrum is characteristic of the substance. By comparing the absorption spectra of various substances with the Fraunhofer lines in the solar spectrum, the gases of the sun's atmosphere have been identified.

11.13 COLOUR OF THE SKY

The colour of the sky as seen from the earth is due to the scattering of sunlight by the molecules of the earth's atmosphere. The amount of scattering is inversely proportional to the fourth power of the wavelength. This is known as *Rayleigh's scattering law*. Hence the shorter wavelengths are scattered much more than the longer wavelengths. Now the predominant wavelength of the sunlight in the shorter wavelength region of the solar spectrum is that corresponding to the blue colour. This explains why the sky appears blue.

Large particles like water droplets and dust present in the atmosphere do not obey Rayleigh's scattering law. They scatter all colours almost equally. This explains why clouds appear white.

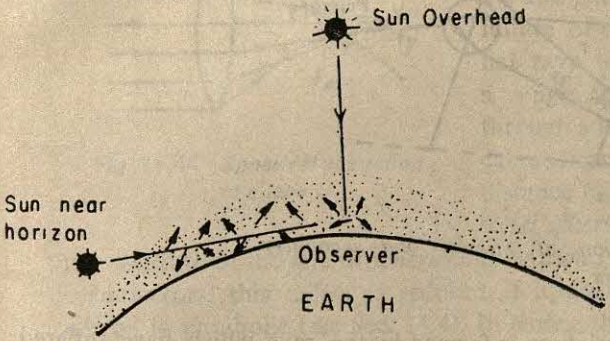


Fig. 11.42 *Absorption of sunlight at sunset and sunrise*

At sunset or sunrise, the rays from the sun have to pass through a greater extent of the atmosphere than at noon. More of the blue colour is scattered away and does not reach the observer (Fig. 11.42). The sun looks almost reddish at sunset and sunrise.

Rainbow

The rainbow is the most beautiful example of the dispersion of white light and production of solar spectrum in nature. When rays from the sun fall on raindrops during or after a shower, they disperse by refraction through water droplets (which act as small prisms) and deviate component colours of the sunlight by internal reflection to the eye of an observer. Sometimes two rainbows are seen. The inner rainbow (which is brighter) is called the *primary* rainbow and the outer rainbow is called the *secondary* rainbow.

The primary rainbow is formed by the rays which have suffered one internal reflection before finally emerging from the water droplets. The red rays emerge from the droplets at an angle of about 43° and the violet rays emerge at an angle of 41° (Fig. 11.43). It exhibits all the colours of the solar spectrum, red being on the outside and violet on the inside.

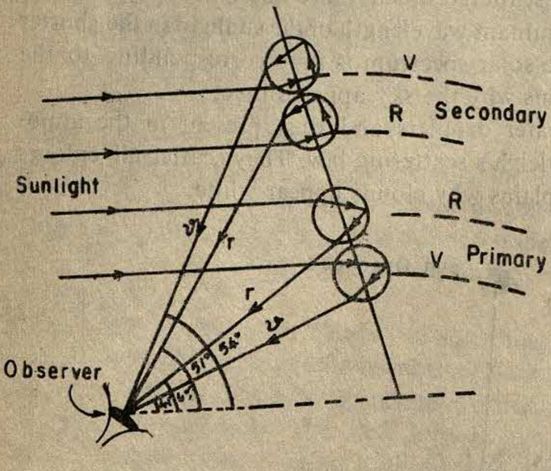


Fig. 11.43 *Formation of primary and secondary rainbows*

The secondary rainbow is formed by the rays which have suffered two internal reflections before finally emerging from the water droplets, as shown in Fig. 11.43. It exhibits the same succession of colours as in the primary rainbow, but in the *reverse* order; red being on the inner edge and violet on the outer edge. The red rays emerge from the water droplets at an angle of 51° and the violet rays at an angle of 54°.

11.14 DEFECTS IN OPTICAL IMAGES

The simple formulae of mirrors and lenses are based on the assumption that the aperture is small and that the object is very close to the principal axis so that the rays are paraxial. These assumptions do not hold good in the case of optical instruments such as telescopes and microscopes which are designed to produce bright and magnified images and for this purpose the objectives of large apertures are necessary. In such cases the behaviour of mirrors and lenses deviates from that given by the simple formulae. These deviations were first investigated by Seidel in 1855.

The optical defects are of two kinds: (i) *monochromatic aberrations* (i.e. defects of images produced by light of a single colour or wavelength) and *chromatic aberration* (i.e. defect of images formed by light of composite colours).

1. Monochromatic Aberrations

There are four common monochromatic aberrations. They are as follows.

(i) *Spherical Aberration* In Sec. 11.4 we have discussed spherical aberration in spherical mirrors. Figure 11.44 shows spherical aberration in a convex lens. Notice that the marginal rays come to a focus F_m and paraxial rays to F_p. This failure of the paraxial and marginal rays to come to a focus at a single point after refraction through a lens is called the spherical aberration of the lens. The distance F_mF_p is a measure of the *axial spherical aberration* and XY is a measure of the *lateral spherical aberration.*

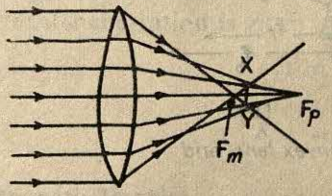


Fig. 11.44 *Spherical aberration in a lens*

In mirrors, this defect is reduced if aperature is small or if the mirror is parabolic (see Sec. 11.4). In lenses, this defect is reduced by decreasing the aperture by putting a stop in front of marginal rays or by using plano-convex lenses.

(ii) *Coma* The image of an object situated just off the axis of a lens has a comet-like appearance; hence the name coma. It is due to the fact that different circular zones of the lens are not symmetrical with respect to points off or away from the axis of the lens.

(iii) *Astigmatism* The horizontal and vertical sections are focussed in different planes giving rise to a defect called astigmatism.

(iv) *Distortion* Distortion is a defect of the image which is due to non-uniform magnifications of the image at different points.

2. Chromatic Aberration

A lens can be regarded as a combination of prisms (see Fig. 11.45). Therefore, a lens splits (disperses) a composite light into different colour components just as a prism does. We have seen that the deviation of blue light is greater than that of the red light; the deviations being towards the base of the prism. Thus if a parallel beam of light falls on a lens, the blue and the red components, for example, will be focussed at different points, as shown in Figs 11.45(a) and (b). Thus a lens will produce coloured images of objects. This defect is called chromatic aberration. The difference $\delta f = f_r - f_b$ between the focal lengths of red and blue colours (or any two colours) is a measure of the chromatic aberration in a lens.

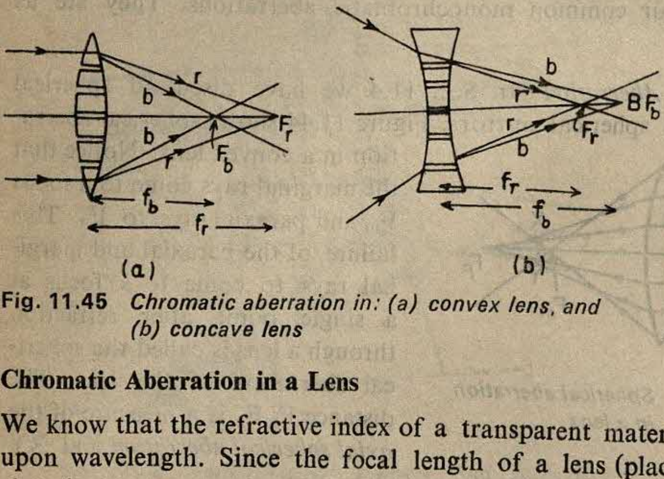


Fig. 11.45 *Chromatic aberration in: (a) convex lens, and (b) concave lens*

Chromatic Aberration in a Lens

We know that the refractive index of a transparent material depends upon wavelength. Since the focal length of a lens (placed in air) is given by

$$\frac{1}{f} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \tag{11.23}$$

it is clear that the focal length will also be different for different wavelengths (or colours). Suppose the light consists of two wavelengths. Let δn represent the change in the refractive index of the lens for the two wavelengths and let δf be the corresponding change in the focal length of the lens. The magnitude $|\delta f|$ is called the chromatic aberration of the lens. The change δf is easily calculated if we use differential calculus.

Differentiating Eq. (11.23), we get

$$-\frac{\delta f}{f^2} = \delta n\left(\frac{1}{R_1} - \frac{1}{R_2}\right)$$

where f is the mean focal length corresponding to the mean refractive index n. Thus

$$-\frac{\delta f}{f} = f\delta n\left(\frac{1}{R_1} - \frac{1}{R_2}\right)$$

But from Eq. (11.23)

$$f\left(\frac{1}{R_1} - \frac{1}{R_2}\right) = \frac{1}{n-1}$$

Therefore

$$-\frac{\delta f}{f} = \frac{\delta n}{n-1} \tag{11.24}$$

Now, we know from Eq. (11.22) that the dispersive power of the material of the lens is given by

$$\omega = \frac{\text{change in } n}{(\text{mean } n) - 1}$$

or

$$\omega = \frac{\delta n}{n-1}$$

Using this definition of ω in Eq. (11.24), the magnitude of the chromatic aberration is given by

$$\delta f = -\omega f$$

or

$$|\,\delta f\,| = \omega f \tag{11.25}$$

Achromatic Doublet

Notice from Figs 11.45(a) and (b) that the sign of δf is opposite in the case of a convex and a concave lens. Hence, a convex lens can be combined with a concave lens such that their chromatic aberrations cancel each other; then the combination will be *achromatic*, i.e. free from chromatic aberration. Such a combination is called an *achromatic doublet* (Fig. 11.46).

Condition for Achromatism Let f and f' be the focal lengths of the convex and concave lenses and let ω and ω' be the dispersive powers of their materials respectively. Then the focal length F of their combination is given by

$$\frac{1}{F} = \frac{1}{f} + \frac{1}{f'}$$

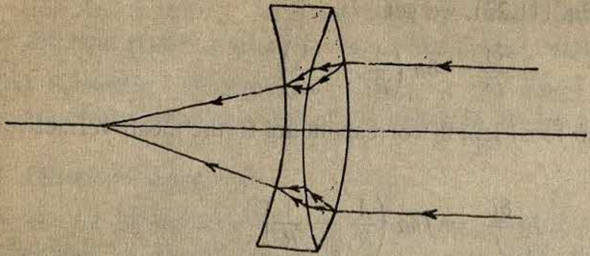


Fig. 11.46 *An achromatic doublet*

Differentiating, we get

$$-\frac{\delta F}{F^2} = -\frac{\delta f}{f^2} - \frac{\delta f'}{f'^2}$$

From Eq. (11.25), we have

$$\frac{\delta f}{f} = -\omega \quad \text{and} \quad \frac{\delta f'}{f'} = -\omega'$$

so that

$$-\frac{\delta F}{F^2} = \frac{\omega}{f} + \frac{\omega'}{f'}$$

Now, if the combination is to be free from chromatic aberration, the focal length F of the combination must not change with colour of light, i.e. $\delta F = 0$ for achromatism. Hence the condition for achromatism is

$$\frac{\omega}{f} + \frac{\omega'}{f'} = 0$$

or

$$\frac{f}{f'} = -\frac{\omega}{\omega'} \qquad (11.26)$$

The negative sign in Eq. (11.26) is due to our sign convention according to which f is positive (convex) and f' is negative (concave). The values of f and f' are related as (without sign convention).

$$\frac{f}{f'} = \frac{\omega}{\omega'}$$

EXAMPLE 11.19 A flint glass convex lens of focal length 20 cm is placed in contact with a crown glass lens. Crown glass has

$$n_b = 1.523 \quad \text{and} \quad n_r = 1.515$$

and flint glass has

$$n'_b = 1.664 \quad \text{and} \quad n'_r = 1.644.$$

(a) Find the nature and focal length of the crown glass lens which will form an achromatic combination with the flint glass lens.

(b) What is the focal length of the combination?

Solution: For crown glass, mean refractive index is

$$n = \frac{1}{2}(n_b + n_r) = 1.519$$

$$\therefore \qquad \omega = \frac{n_b - n_r}{n - 1} = 0.0154$$

For flint glass, mean refractive index is

$$n' = \frac{1}{2}(n'_b + n'_r) = 1.654$$

$$\therefore \qquad \omega' = \frac{n'_b - n'_r}{n' - 1} = 0.0306$$

(a) Let f' be the focal length of the flint glass convex lens and f be the focal length of the crown glass lens. For achromatism,

$$\frac{f}{f'} = -\frac{\omega}{\omega'}$$

where $f' = +20$ cm (convex lens)

$$\therefore \qquad f = -f'\frac{\omega}{\omega'} = -20 \times \frac{0.0154}{0.0306} = -10.1 \text{ cm}$$

The negative sign indicates that the crown glass lens must be concave.

(b) The focal length of the combination is given by

$$\frac{1}{F} = \frac{1}{f} + \frac{1}{f'}$$

$$\therefore \qquad F = -20.4 \text{ cm}$$

EXAMPLE 11.20 Two glasses have dispersive powers in the ratio of 2 : 3. These glasses are used in the manufacture of an achromatic objective of focal length 20 cm. What are the focal lengths of the two lenses of the objective?

Solution: For an achromatic combination,

$$\frac{f}{f'} = -\frac{\omega}{\omega'}$$

$$\therefore \qquad \frac{1}{f'} = -\frac{\omega}{\omega'} \cdot \frac{1}{f}$$

where $\omega/\omega' = 2/3$. Therefore,

$$\frac{1}{f'} = -\frac{2}{3f} \qquad \text{(i)}$$

The focal length of the combination is

$$\frac{1}{F} = \frac{1}{f} + \frac{1}{f'}$$

where $F = +20$ cm. Therefore,

$$\frac{1}{20} = \frac{1}{f} + \frac{1}{f'} \qquad \text{(ii)}$$

From Eqs (i) and (ii), we get

$$f = 6.67 \text{ cm} \qquad \text{and} \qquad f' = -\frac{3f}{2} = -10 \text{ cm}$$

11.15 CAMERA

A photographic camera essentially consists of a converging lens system at one end of a light-proof box and a sensitive photographic film at the other end (Fig. 11.47). The lens system forms a real, inverted and usually diminished image on the film. The lenses are mounted on a slide so that the distance from the film can be altered in order to bring into sharp focus the images of the objects at varying distances. This is called "focussing". An adjustable stop (called aperture) is used to control the amount of light entering the camera. A shutter is placed between the lens system and the film. When a photograph is taken, the shutter opens for a very short time and closes quickly.

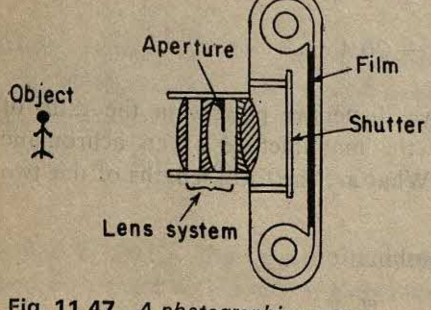


Fig. 11.47 *A photographic camera*

To obtain a good picture, a photographer must control the following factors.

(i) *Exposure Time* An important parameter is the time for which the film should be exposed so as to obtain a clear image. If the object to be photographed is in broad daylight, the exposure time should be small. Under cloudy conditions, in the shade or indoors, the exposure time has to be increased. Some normal exposure times are

$$\frac{1}{500}\text{ s}, \frac{1}{250}\text{ s}, \frac{1}{125}\text{ s}, \frac{1}{60}\text{ s}, \frac{1}{30}\text{ s, etc.}$$

(ii) *The f-number* The amount of light entering the camera is controlled by means of the aperture which has circular opening whose diameter is adjustable. The size of the circular opening is expressed in terms of the so-called *f*-number which is defined as follows:

$$f\text{-number} = \frac{\text{focal length of the lens}}{\text{diameter of the aperture}}$$

Camera lenses are usually specified by two numbers: the focal length and the *f*-number. Thus the term "8 cm, *f*/16 lens" means that the focal length of the lens is 8 cm and its *f*-number $= f/16 = 8/16 = 0.5$ cm, i.e. the diameter of the aperture is 5 mm. Some typical apertures used in a camera are:

$$\frac{f}{2}, \frac{f}{2.8}, \frac{f}{4}, \frac{f}{5.6}, \frac{f}{8}, \frac{f}{11}, \frac{f}{16}, \text{ etc.}$$

(iii) *Film Speed* Films can be "fast" or "slow". A fast film requires a relatively short time to be exposed, and is therefore used in poor-light conditions. Slow films require a longer time and are thus suitable for still photography.

(iv) *Depth of Focus* Under poor lighting conditions, the aperture has to be made large. With a large aperture, the image is not focussed at the focal plane but gets spread out owing to aberrations. The depth of focus indicates the range of object distances over which good focussing is achieved.

11.16 MICROSCOPE

In the simplest form, a simple microscope or the magnifying glass is just a thin convex lens. The object to be examined is placed between the centre and the focus of the lens and the eye is placed just behind the lens. An enlarged, erect, virtual image of the object is formed by the lens (Fig. 11.48). A convex lens used in this way is called a magnifying glass or a simple microscope. Because of the fact that a normal eye is able to see the details of an object most distinctly when it is at a distance of $D = 25$ cm approximately, (the distance of distinct vision) from the eye, the position of the lens is usually adjusted so that the virtual image is formed at this distance from the eye.

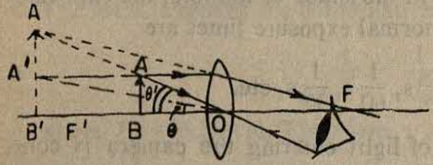


Fig. 11.48 *The simple microscope*

The magnifying power of an optical instrument is defined as

$$M = \frac{\theta'}{\theta} = \frac{\text{Angle subtended by the image formed at the distance of distinct vision}}{\text{Angle subtended by the object when placed at the distance of distinct vision}}$$

$$= \frac{\angle \text{A'OB'}}{\angle \text{A''OB}}$$

Since the eye is held very close to the lens, the angles subtended at the eye may be taken to be the same as the angles subtended at the lens. The object AB to be examined by a microscope is extremely tiny. So is the image (the sizes of the object and the image are shown exaggerated in the diagram). Therefore, angles θ and θ' are extremely small. For small angles (expressed in radians)

$$\sin\theta \simeq \theta \qquad \cos\theta \simeq 1 \qquad \tan\theta \simeq \theta$$

$$\therefore \qquad M = \frac{\tan\theta'}{\tan\theta} = \frac{\text{AB/OB}}{\text{A''B'/OB'}} = \frac{\text{OB}'}{\text{OB}}$$

or
$$M = \frac{-D}{-x} = \frac{D}{x}$$

Now putting $v = -D$, $u = -x$ in $\frac{1}{v} - \frac{1}{u} = \frac{1}{f}$ gives

$$\frac{D}{x} = 1 + \frac{D}{f}$$

Therefore
$$M = 1 + \frac{D}{f} \tag{11.27}$$

Thus, the shorter the focal length, the larger is the magnifying power.

The Compound Microscope

With a simple microscope, a magnification of about 20 can be achieved. When higher magnifications are desired, a compound microscope is used. It was invented by Galileo in 1610. It consists of two converging lenses, called the *objective* and the *eyepiece*. The objective is a convex lens of very short focal length and is placed close to the object. The eyepiece is also a convex lens of slightly longer focal length and is placed close to the eye.

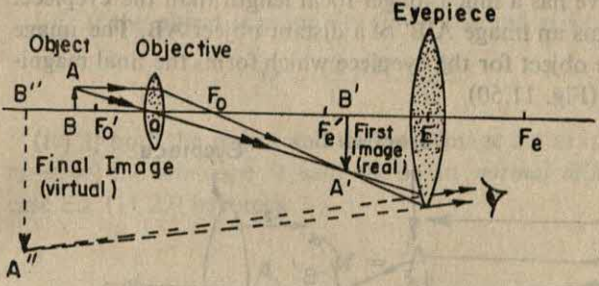


Fig. 11.49 *The compound microscope*

The object AB to be examined is placed just beyond the principal focus F_0' of the objective as shown in Fig. 11.49, forming a slightly enlarged real inverted image A′B′. This image is merely formed in space, the light going on to meet the eyepiece which examines this image acting like a simple magnifying glass. The eye piece is moved up until A′B′ lies just within its focus F_e'. Consequently, a highly enlarged virtual image A″ B″ is produced, the real image A′B′ acting as the object for the eyepiece. This final image is seen by the eye.

The magnifying power of the microscope is equal to the product of the magnifying powers of the objective and the eyepiece, i.e.

$$M = M_0 \times M_e$$

$$= \left(1 + \frac{v}{f_0}\right)\left(1 + \frac{D}{f_e}\right)$$

where $D = 25$ cm and v = distance of the final image from the objective. The magnifying power of the objective is approximately equal to L/f_0 where L = length of the microscope tube (= distance between the objective and the eyepiece) and that of the eyepiece is approximately equal to D/f_e so that

$$M \simeq \frac{L}{f_0} \cdot \frac{D}{f_e} \tag{11.28}$$

The objective and eyepiece are not single lenses but combinations of lenses so that the defects in the final image are minimized.

11.17 TELESCOPE

A telescope is an optical instrument that is used to examine distant objects. Two types of telescopes are in use—refracting and reflecting telescopes.

Refracting Telescope

It consists of two convergent lenses called the objective and the eyepiece. The objective has a much longer focal length than the eyepiece. The objective forms an image A′B′ of a distant object AB. The image A′B′ serves as the object for the eyepiece which forms the final magnified image A″B″ (Fig. 11.50).

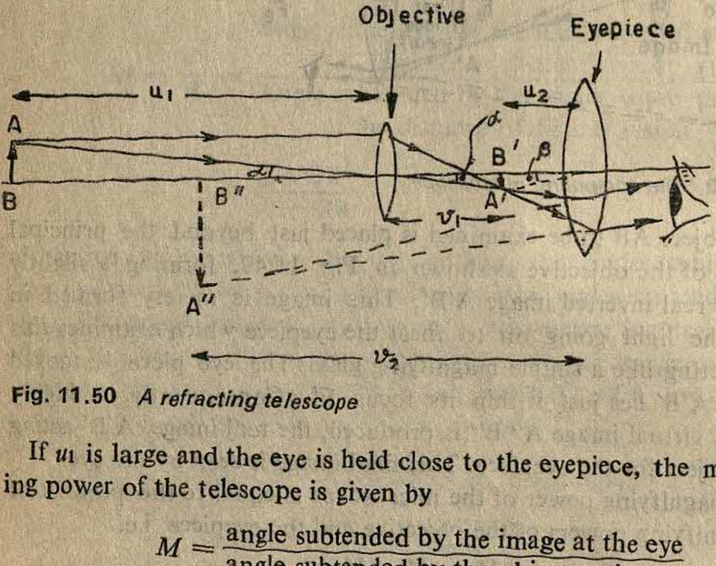


Fig. 11.50 *A refracting telescope*

If u_1 is large and the eye is held close to the eyepiece, the magnifying power of the telescope is given by

$$M = \frac{\text{angle subtended by the image at the eye}}{\text{angle subtended by the object at the eye}}$$

$$= \frac{\beta}{\alpha} = \frac{A'B'/u_2}{A'B'/v_1} = \frac{v_1}{u_2}$$

Now $\frac{1}{v_1} - \frac{1}{u_1} = \frac{1}{f_0}$ and $\frac{1}{v_2} - \frac{1}{u_2} = \frac{1}{f_e}$, where f_0 and f_e are the focal lengths of the objective and the eyepiece respectively. Hence

$$M = \frac{u_1 f_0}{(u_1 + f_0)} \cdot \frac{(v_2 + f_e)}{v_2 f_e} \tag{11.29}$$

where u_1 = distance of the object from the objective
and v_2 = distance of the final image from the eyepiece

Special Cases: (i) If the final image is formed at the distance of distinct vision, $v_2 = D$, then

$$M = \frac{u_1 f_0}{(u_1 + f_0)} \cdot \frac{(f_e + D)}{f_e D}$$

(ii) If the final image is formed at infinity, $v_2 \rightarrow \infty$, then Eq. (11.29) reduces to

$$M = \frac{u_1}{(u_1 + f_0)} \cdot \frac{f_0}{f_e}$$

(iii) If the object is at infinity ($u_1 \rightarrow \infty$) and $v_2 = D$, then

$$M = \frac{f_0}{f_e} \cdot \frac{(D + f_e)}{D}$$

(iv) If both the object and the final image are at infinity ($u_1 \rightarrow \infty$, $v_2 \rightarrow \infty$) the telescope is said to be in *normal adjustment*, in which case Eq. (11.29) becomes

$$M = \frac{f_0}{f_e}$$

From this expression, it is clear that for high magnification, $f_0 \gg f_e$, i.e. the objective must have a long focal length and the eyepiece a very short focal length. The telescope at Yerke's observatory has $f_0 \simeq 19.3$ m and $f_e \simeq 0.51$ cm, giving a magnification of about 4000.

If the telescope is to be used for heavenly bodies, it does not matter if the final image is inverted. But if it is to be used to see distant objects on earth, the final image must be erect. For this purpose an additional lens, called the *erecting lens* is used between the objective and the eyepiece which are achromatic combinations.

Reflecting Telescope

Most of the large astronomical telescopes of the world today employ a concave parabolic mirror in place of the achromatic objective. The main advantage is that a parabolic mirror is free from chromatic and spherical aberrations, thus permitting a large aperture and hence giving a high resolving power (see Sec. 11.18). These are of great importance in the study of stars which are close together.

Figure 11.51 shows the optical features of the great 200-inch aperture reflecting telescope. It is of the Cassegrainian type. With the large concave mirror C parallel rays entering the telescope would have been brought into focus at F. But before reaching F the converging bundle of rays is reflected by a convex mirror m down the axis and passing through a hole in the objective, comes to focus at S, a point

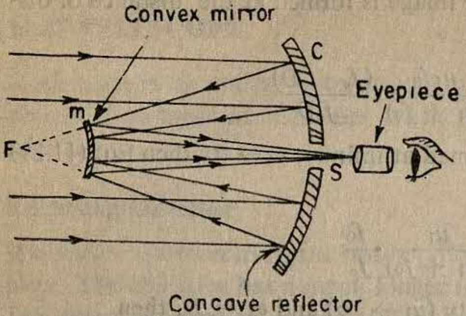


Fig. 11.51 *Reflecting telescope, Cassegrain type*

just outside the hole so that F and S are conjugate foci for the convex mirror m. For visual observations the image S falls in the first focal plane of the eyepiece, so that the final image is at infinity. This image is inverted since the second image at S is erect with respect to the first image at F and this first image is inverted with respect to the object.

11.18 RESOLVING POWER

So far we have not considered what effect the wave nature of light will have on the images formed in optical instruments. In the previous chapter we have seen that if a narrow slit is illuminated, it produces a diffraction pattern consisting of alternate bright and dark straight fringes. If the slit is replaced by a circular aperture (a small hole), the diffraction pattern consists of alternate bright and dark circular fringes. Now suppose a lens is used to form an image of an object. The lens itself can be regarded as a circular aperture which forms a diffraction pattern. The image of each point is a set of circular fringes with a bright patch at the centre. The size of this patch depends upon the aperture of the lens and the wavelength of light.

Suppose we have two pin-holes on a piece of paper and illuminate them and look at their images formed by a lens. If the holes are far apart, their diffraction patterns will not overlap and the two images are clearly visible. If the holes are very close together, their diffraction patterns will overlap. Consequently, their images can no longer be identified. The resolving power of a lens is the ability to resolve the two images so that they are distinctly identified.

Resolving Power of Human Eye

The human eye has an aperture (called pupil). The diameter of the aperture is about 2 mm in normal light. Two illuminated pin-holes close together can be seen distinctly (as separate pin-holes) if they subtend at the eye an angle of about one minute of arc. The reciprocal of this angle gives the resolving power of the eye.

Resolving Power of a Microscope

The resolving power of a microscope is determined by the smallest distance between two point objects (such as pin-holes) which can be distinguished by it. This distance d is given by

$$d = \frac{\lambda}{2n \sin \theta}$$

where λ = the wavelength of light used to illuminate the objects, θ = half the angle of the cone of light from each point object and n = refractive index of the transparent medium between the objects and the objective of the microscope. Normally the medium is air ($n = 1$). The resolving power of the microscope is the reciprocal of the distance d. The smaller the value of d, the greater will be the resolving power. An increase in n will produce a greater resolving power. It is for this reason that the objective is immersed in an oil (called the oil-immersion objective) and the object to be examined placed in oil.

Resolving Power of a Telescope

The resolving power of a telescope is defined as the reciprocal of the smallest angular separation between two distant objects whose images are distinctly separated by the telescope. This angle is given by

$$\delta\theta = \frac{1.22\,\lambda}{a}$$

where λ is the wavelength of light and a is the diameter of the telescope objective (i.e. its aperture). The higher the aperture, the smaller will be the value of $\delta\theta$ and, consequently, the greater will be the resolving power. It is for this reason that the aperture of the objective is made very large (~100 cm or more). Such a telescope will clearly resolve two stars very close together, called binary stars.

SUMMARY

The light from a source, and received on a surface, is measured in terms of luminous intensity, luminous flux and illuminance. The speed of light in free space is 3.0×10^8 ms^{-1}.

Spherical mirrors and lenses form images of illuminated or luminous objects. They are employed in optical instruments such as cameras, microscopes and telescopes. These images suffer from optical defects such as spherical aberration, coma, distortion, astigmation and chromatic aberration. Combinations of lenses are used to minimize some of these defects.

The resolving power of an optical instrument is its ability to distinguish two nearby point objects.

EXERCISES

A. Short-Answer Questions

1. State with reason whether the following statements are *true* or *false*.
   - (i) Illuminance obeys the inverse square law of distance.
   - (ii) A concave mirror forms only real images.
   - (iii) A convex mirror forms only virtual diminished images.
   - (iv) A camera will not work if its convex lens is replaced by a concave lens.
   - (v) The focal length of a spherical mirror is always half its radius of curvature.
   - (vi) For paraxial rays, the focal length of an equi-convex lens placed in air is equal to the radius of curvature of either face, provided the the aperture is small and the lens is thin.
   - (vii) The critical angle for total internal reflection for glass of refractive index $\sqrt{2}$ is 45°.
   - (viii) At minimum deviation, the refracted ray in the prism subtends equal angles with the refracting faces of the prism.
   - (ix) The refractive index of the material of a prism depends upon the angle of the prism and the angle of minimum deviation.
   - (x) A crown glass prism has a higher dispersive power than a flint glass prism.

B. Long-Answer Questions

1. Define the following terms and state and define their SI units: luminous intensity, luminous flux and illuminance.
2. The efficiency of a source is 50 lmW^{-1}. What is the meaning of this statement?
3. Describe: (a) Fizeau's method and (b) Michelson's method, for the measurement of the speed of light.
4. Stating the sign convention and the principle of image formation, deduce the relation between the focal length and the object and image distances in the case of: (a) a concave mirror and (b) a convex mirror.
5. What do you understand by the term "linear magnification?" Stating the sign convention, obtain the expression for the linear magnification of the image in terms of the object and image distances in the case of: (a) a concave mirror and (b) a convex mirror.

6. Stating the sign convention and the principle of image formation, deduce the relation between the focal length and the object and image distances in the case of: (a) a convex lens and (b) a concave lens.
7. What is meant by the linear magnification" of the image? Stating the sign convention, obtain the expression for the linear magnification in terms of the object and image distances for: (a) a convex lens and (b) a concave lens.
8. Discuss, by drawing ray diagrams or by using the lens formula, the nature, position and size of the image formed by a convex lens of focal length f when the object is placed: (a) at infinity, (b) beyond $2f$, (c) between f and $2f$, (d) at f and (e) between the focus and the optical centre.
9. Describe an optical experiment to distinguish between a convex lens, a concave lens and a thin plane glass plate.
10. (a) What is total internal reflection? State the conditions necessary for this phenomenon.

    (b) What is the critical angle? How it is related to refractive index?
11. What is a mirage? How is it explained?
12. What are optical fibres? How do they transmit light? State two applications of optical fibres.
13. Show that the refractive index of the material of a prism is given by

$$n = \frac{\sin\left[\frac{1}{2}\ (A + \delta_m)\right]}{\sin\left(\frac{A}{2}\right)}$$

    where the symbols have their usual meanings.
14. Deduce the relation between the object distance, image distance and the radius of curvature of: (a) a convex refracting surface and (b) a concave refracting surface.
15. Deduce the lens makers's formula. Clearly state the limitations of the formula.
16. Two lenses of focal lengths f_1 and f_2 are placed in contact coaxially. Obtain the focal length of the combination.
17. What is dispersion of light? How will you demonstrate it experimentally?.
18. Describe and explain the formation of a rainbow.
19. Describe the function of the essential parts of a spectroscope. How is it used to determine the dispersive power of a transparent substance such as glass?
20. What is a direct vision spectroscope? How is it a useful device.
21. What is an emission spectrum? Distinguish between continuous, line and band spectra, giving one example of each.
22. What do you understand by the absorption spectrum of a substance? Describe an experiment to observe the absorption spectrum of sodium.
23. What are Fraunhofer lines in the solar spectrum? What are they due to?
24. Give a brief account of the various monochromatic defects of images.
25. What is chromatic aberration? Deduce the condition of the achromatism of a combination of two lenses.
26. Name the essential parts of a camera. What are the factors that must be borne in mind while using a camera?

27. Draw a ray diagram to illustrate action of a compound microscope. On what factors does its magnifying power depend?
28. Describe the action of a refracting telescope and deduce the expression for its magnifying power when used in normal adjustment.
29. What is a reflecting telescope? What is its main merit over a refracting telescope?
30. What do you understand by resolving power. Define the resolving power of: (a) the human eye, (b) a microscope and (c) a telescope.

C. Multiple-Choice Questions

Choose the correct answer from the given alternatives.

1. Which of the following parameters does not change when light travels from air into a glass slab?

   (a) Velocity
   (b) Wavelength
   (c) Frequency
   (d) Amplitude

2. The refractive index of a glass prism depends upon

   (a) the angle of the prism
   (b) the angle through which it deviates an incident beam of light
   (c) the colour of the incident light
   (d) the intensity of the incident light

3. A prism of refractive index of $\sqrt{2}$ has a refracting angle of 60°. At what angle must a ray be incident on it so that it suffers a minimum deviation?

   (a) 30°
   (b) 45°
   (c) 60°
   (d) 75°

4. Two transparent media A and B are separated by a plane boundary. The speed of light in medium A is 2.0×10^8 ms^{-1} and in medium B 2.5×10^8 ms^{-1}. The critical angle for which a ray of light going from A to B is totally internally reflected is

   (a) $\sin^{-1}(1/2)$
   (b) $\sin^{-1}(2/5)$
   (c) $\sin^{-1}(4/5)$
   (d) $\sin^{-1}(4/7)$

5. A convex glass lens ($n_g = 1.5$) has a focal length of 8 cm when placed in air. What is the focal length of the lens when it is immersed in water ($n_w = 4/3$)?

   (a) 4 cm
   (b) 8 cm
   (c) 16 cm
   (d) 32 cm

6. A plano-concave lens is made of glass of refractive index 1.5 and the radius of curvature of the curved face is 50 cm. The power of the lens is

   (a) -1.0 D
   (b) -0.5 D
   (c) $+1.0$ D
   (d) $+0.5$ D

7. A person can see clearly only up to a distance of 30 cm. He wants to read a book placed at a distance of 50 cm from his eyes. What is the power of the lens he requires for his spectacles?

   (a) -1.0 D
   (b) -1.33 D
   (c) -1.67 D
   (d) -2.0 D

8. A person with defective eyesight is unable to see objects clearly nearer than 1.5 m. He wants to read a book placed at a distance of 30 cm from his eyes. Find the power of the lens he requires.

   (a) +2.67 D (b) −2.67 D
   (c) +2.5 D (d) −2.5 D

9. The principal section of a glass prism ($n = 1.5$) is an equilateral triangle. A ray is incident normally on one of the faces. The angle between this ray and the ray that leaves the prism is

   (a) 60° (b) 90°
   (c) 120° (d) 180°

10. A plano-convex lens is made of glass of refractive index 1.5. The focal length f of the lens and radius of curvature R of its curved face are related as

    (a) $f = R/2$ (b) $f = R$
    (c) $f = 2\,R$ (d) $f = 3R/2$

11. A thin convergent glass lens ($n_g = 1.5$) has a power of +5.0 D. When this lens is immersed in a liquid of refractive index n_l it acts as a divergent lens of focal length 100 cm. The value of n_l must be

    (a) 4/3 (b) 5/3
    (c) 5/4 (d) 6/5

12. The dispersion of light in a medium implies that

    (a) lights of different wavelengths travel with different speeds in the medium
    (b) lights of different frequencies travel with different speeds in the medium
    (c) the refractive index of the medium is different for different wavelengths of light
    (d) all of the above

13. When a telescope is in normal adjustment, the distance of the objective from the eyepiece is found to be 100 cm. If the magnifying power of the telescope, at normal adjustment, is 24, the focal lengths of the lenses are

    (a) 96 cm, 4 cm (b) 90 cm, 10 cm
    (c) 80 cm, 20 cm (d) 50 cm, 50 cm

14. Monochromatic light is refracted from air into glass of refractive index n. The ratio of the wavelengths of the incident and refracted waves is

    (a) 1 : 1 (b) $1 : n$
    (c) $n : 1$ (d) $n^2 : 1$

D. Numerical Problems

1. Two plane mirrors are inclined to each other at an angle θ. A ray incident on one mirror at 50° after reflection from it falls on the second mirror where it is reflected parallel to the first. Determine the value of θ.
2. A narrow beam of light after reflection from a plane mirror falls on a scale 100 cm from the mirror. When the mirror is rotated a little, the spot moves through 20 cm. Determine the angle through which the mirror is rotated.
3. A lamp is 4 m from a wall. Find the focal length of a concave mirror which will form a five times magnified image of the lamp on the wall. How far from the wall must the mirror be placed?

4. A dentist's concave mirror has a radius of curvature of 30 cm. How far must it be placed from a small cavity in order to give a virtual image magnified five times?
5. When an object is moved 12 cm towards a concave mirror, the real image moves out from 25 to 75 cm. Find the radius of curvature of the mirror.
6. An object is placed 25 cm from a convex mirror and a plane mirror is placed at a distance of 20 cm from the object so that the images formed in the two mirrors coincide without parallax. Determine the focal length of the convex mirror.
7. When an object is kept at a distance of 30 cm from a convex mirror, the image is formed at a distance of 10 cm from the mirror. If the object is moved with a speed of 9 ms^{-1}, find the speed of the image.
8. A coin 2.54 cm in diameter held 254 cm from the eye just covers the full disc of the moon. What is the angular diameter of the image of the moon formed by a concave mriror of radius of curvature 1.27 m?
9. A convex mirror produces a magnification of 1/2 when the object is at a position A and a magnification of 1/3 when the object is at a position B. If positions A and B are 60 cm apart, find the focal length of the mirror.
10. A rectangular block of glass is placed on a printed page lying on a horizontal surface. Find the minimum value of the refractive index of glass for which the letters on the page are not visible from any of the vertical sides of the block.
11. A prism of refracting angle 60° produces a deviation of 40° for two angles of incidence which differ by 11°. Determine the refractive index of the material of the prism.
12. The refracting angle of a prism is A and it produces a minimum deviation of $(180° - 2A)$. Show that the refractive index of the prism is given by

$$n = \frac{1}{\tan(A/2)}$$

13. A ray of light falls at grazing incidence on one of the faces of a prism of refractive index n and refracting angle A. Show that the angle of emergence is given by

$$\sin e = (n^2 - 1)^{1/2} \sin A - \cos A$$

14. The two surfaces of a concave lens are of radii of curvature 10 cm and 30 cm. If the refractive index of the lens is 1.5, find its focal length when it is immersed in water ($n_w = 4/3$).
15. An object and a screen are set up 80 cm apart. There are two positions at which a convex lens forms a real image on the screen, the magnifications in the two cases being 2/3 and 3/2. Find the focal length of the lens.
16. A plano-convex lens of focal length 15 cm, is fixed on the side of a box with the convex surface outside. An object 5 cm high is placed at 25 cm, from the lens. Find the position and size of the image when the box is (a) empty and (b) filled with water. Refractive index of lens = 1.5 and that of water = 4/3.
17. A thin equi-convex lens of glass of refractive index 1.5 and radius of curvature of surfaces 24 cm is placed on a horizontal plane mirror. When the space between the lens and the mirror is filled with a liquid, a pin held 40 cm vertically above the lens is found to coincide with its own image. Calculate the refractive index of the liquid.

18. The dispersive powers of crown and flint glasses are 0.015 and 0.030 respectively. Calculate the focal lengths of the lenses (made of crown and flint glass) which form an achromatic doublet of focal length 60 cm when placed in contact.
19. A compound microscope has an objective of focal length 5 cm and an eyepiece of focal length 10 cm separated by a distance of 16 cm. Find the magnification of the image formed at a distance of 25 cm from the eyepiece.
20. An astronomical telescope has an objective of focal length 100 cm and an eyepiece of focal length 20 cm. The diameter of the objective is 5 cm. Find the magnifying power in normal adjustment, the separation between the lenses and the diameter of the image.

12

ELECTRONS AND PHOTONS

12.1 INTRODUCTION

The word "atom" is derived from a Greek word which literally means indivisible. Democritus, the Greek philosopher, considered that all matter consists of indivisible particles called atoms. We now know that this is a misnomer, for any atom can be subdivided. The discovery of electricity motivated research that has led to the present-day concept of an atom. Michael Faraday studied the passage of electricity through liquid solutions. His laws of electrolysis conclusively established that matter is electrical in nature, i.e. matter is composed of positive and negative charges. The next step in the study of the electrical nature of matter was made by J.J. Thomson in 1897. Whereas Faraday studied the passage of electricity through liquids, Thomson focussed his attention on the passage of electricity through gases. The studies of Thomson on the electrical discharge through gases at low pressure established that the atom has a structure, i.e. an atom is composed of charged particles, e.g. electrons. Historically, the study of electrical discharges through gases and the discovery of cathode rays, i.e. electrons, marked the beginning of a new branch of physics called atomic physics.

12.2 ELECTRICAL DISCHARGE THROUGH A GAS

The conduction of electricity through gases has been studied in a variety of experiments and has resulted in several important discoveries. The apparatus consists of a long glass tube about 4 cm in diameter and 100 cm in length, having a side tube that is connected to a vacuum pump and a manometer (Fig. 12.1). Two electrodes, a cathode and an anode, are enclosed in the tube. A high voltage of 10,000 to 15,000 V is applied between the electrodes.

The tube is gradually evacuated. The first discharge appears when the pressure of air drops to about 100 mm of Hg [Fig. 12.2(a)]. This

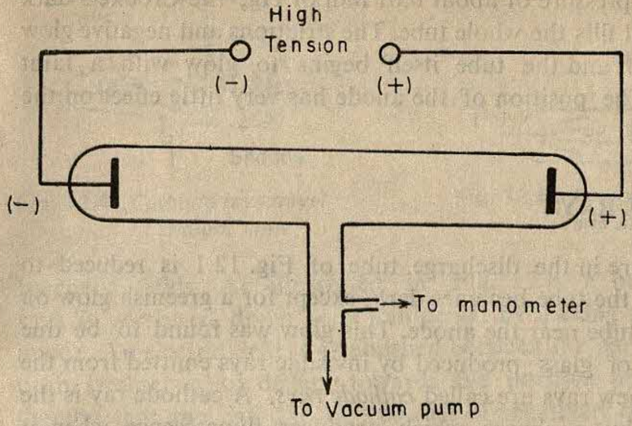


Fig. 12.1 *Experimental set-up to study electrical discharge through air at low pressure*

discharge consists of long, thin lines of irregular form, which are bluish in colour. These lines are called *blue streamers.*

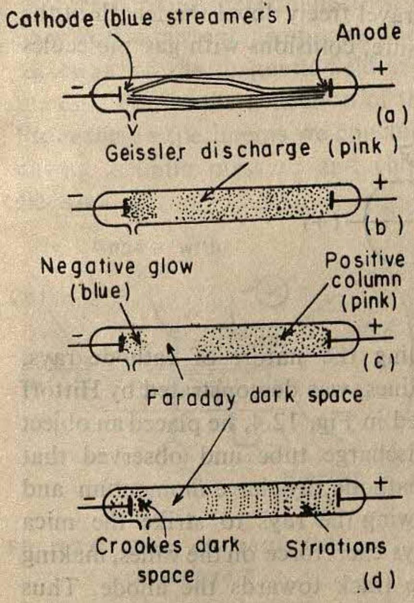


Fig. 12.2 *The general appearance of a high voltage electrical discharge through a gas at various stages of evacuation*

When the gas pressure is reduced to about 5 mm of Hg, the discharge (called *Geissler discharge*) changes to pink and widens to fill the whole tube [Fig. 12.2(b)]. As the pressure is reduced further to about 2 mm of Hg, the bright discharge divides into two parts: a long positive column (pink) and a short blue section (called the *negative glow*) separated by a dark region called the *Faraday dark space* [Fig. 12.2(c)]. At still lower pressures (1 mm of Hg) the Faraday dark space becomes bigger and the negative glow moves away from the cathode and another dark space (called *Crooke's dark space*) [Fig. 12.2 (d)] appears between the cathode and the negative glow. The positive column is divided into equally spaced layers called *striations*. When the tube is evacuated still further to a pressure of about 0.01 mm of Hg, the Crooke's dark space widens and fills the whole tube. The striations and negative glow become very faint and the tube itself begins to glow with a faint greenish light. The position of the anode has very little effect on the discharge.

12.3 CATHODE RAYS

When the pressure in the discharge tube of Fig. 12.1 is reduced to 10^{-4} mm of Hg, the tube becomes dark, except for a greenish glow on the walls of the tube near the anode. This glow was found to be due to fluorescence of glass produced by invisible rays emitted from the cathode. These new rays are called *cathode rays.* A cathode ray is the path traversed by a charge which produces fluorescence when it impinges on the walls of the tube (Fig. 12.3). If a small hole is made in the anode, a green spot is seen on the glass behind the anode. Since the cathode rays travel towards the positive anode, they must carry a negative charge. These rays are the path of tiny particles called *electrons.* In a highly evacuated tube, these particles, having been detached from the atoms of the cathode, travel freely down the length of the tube. Because of extremely low pressure, collisions with gas molecules are very rare.

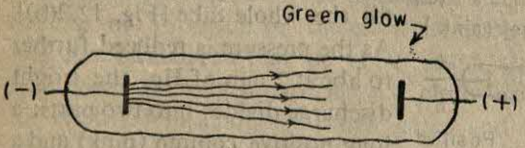


Fig. 12.3 *Cathode rays*

The first important fact regarding the nature of cathode rays, namely, that they travel in straight lines, was demonstrated by Hittorf in 1869. In his experiment, illustrated in Fig. 12.4, he placed an object in the path of the rays inside the discharge tube and observed that they cast its shadow. The fact that cathode rays have momentum and energy can be demonstrated by allowing the rays to strike the mica vanes (Fig. 12.5) of a wheel. The rays exert force on the vanes, making them rotate and roll along a double track towards the anode. Thus we conclude that the cathode particles have momentum and energy and therefore, a mass and velocity like material particles.

The fact that cathode rays are negatively charged particles can be demonstrated by studying the deflection of the rays by electric and

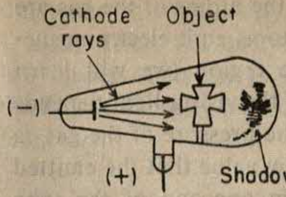


Fig. 12.4 *Cathode rays travel in straight lines*

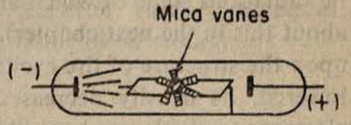


Fig. 12.5 *Cathode rays have energy and momentum*

magnetic fields. As illustrated in Fig. 12.6, a beam of cathode rays is narrowed down to a thin pencil by making a fine hole in the anode. An electric field is established between the parallel plates. The beam is observed to deflect towards the positive plate, establishing thereby that the cathode rays consist of negatively charged particles. On the other hand, such a beam should also be deflected by a magnetic field directed perpendicular to its track. By placing a horseshoe magnet, as shown in Fig. 12.7, the cathode rays are made to bend down. If the polarity of the magnet is reversed, they are bent upwards. The bending shows that the particles are charged and that the direction of bending determines their charge. Streaming charged particles constitute a current. Applying the left-hand rule and knowing the direction of the magnetic field and that of the current, we infer from the direction of the deflection of the beam that the charge is negative. From these experiments we conclude that a cathode ray is a particle having a finite mass m and charge e. This particle is known as the electron.

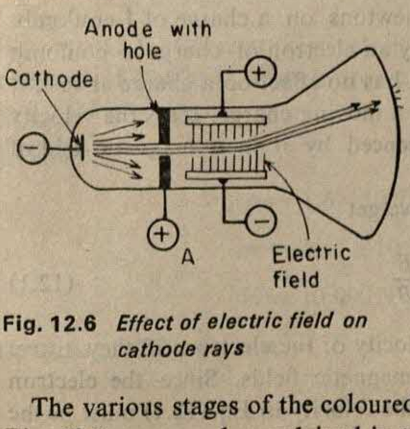


Fig. 12.6 *Effect of electric field on cathode rays*

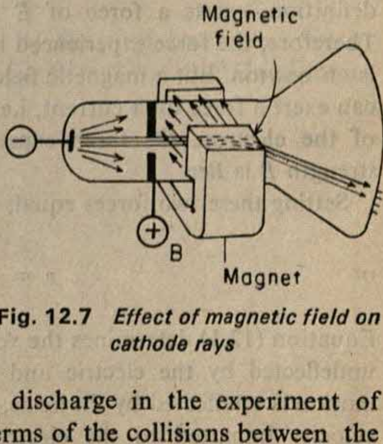


Fig. 12.7 *Effect of magnetic field on cathode rays*

The various stages of the coloured discharge in the experiment of Fig. 12.1 can now be explained in terms of the collisions between the electrons emitted from the cathode and the atoms of the gas in the tube. Owing to collisions with fast electrons, the atoms of the gas are excited to higher energy states. The excited atoms emit electromagnetic radiations some of which are in the visible region (we will learn about this in the next chapter). The colour of the visible light depends upon the structure of the excited atoms. As the pressure of the gas is lowered, its density decreases to such a low value that the emitted electrons (cathode rays) are able to travel from one end of the tube to the other without undergoing collisions with the atoms of the gas. Therefore, the atoms remain in the unexcited state and hence cannot emit electromagnetic radiation. At this stage the tube appears dark.

12.4 CHARGE TO MASS RATIO OF ELECTRONS

By combining the two experiments shown in Figs 12.6 and 12.7, J.J. Thomson was able to determine the charge to mass ratio (*e*/*m*) of an electron. In chapter 5, we have learnt how he modified the discharge tube for this purpose (see Fig. 5.27). We have learnt how he measured the value of *e*/*m*. Here we describe the basic physics of his ingeneous experiment.

He applied electric and magnetic fields at right angles to each other, each at right angles to the beam of electrons and adjusted the strengths of the two fields so that the upward deflection of the beam by the electric field was completely cancelled by the downward deflection by the magnetic field. The beam of electrons continued straight without any deflection. The deflection of a charged particle depends on how much force it experiences in a field. An electric field of strength E, by definition, exerts a force of E newtons on a charge of 1 coulomb. Therefore, the force experienced by an electron of charge e coulomb is eE newton. But a magnetic field has no effect on a charge at rest; it can exert a force on a current, i.e. a moving charge. If v is the velocity of the electron, the force experienced by it in a magnetic field of strength B is Bev.

Setting these two forces equal, we get

or

$$v = \frac{E}{B} \tag{12.1}$$

Equation (12.1) determines the velocity of the electrons if they travel undeflected by the electric and magnetic fields. Since the electron motion is unaffected by the fields, the velocity in Eq. (12.1) must be the velocity imparted to the electrons by the voltage V volts applied between the cathode and the anode. By definition, if the potential difference is 1 volt, then the work done on a charge of 1 coulomb is 1 joule. Thus, the energy gained by an electron of charge e coulomb is eV joules. This energy is gained by the electron which must equal the kinetic energy $\frac{1}{2}mv^2$ acquired by it by virtue of its motion. Hence,

$$eV = \frac{1}{2}mv^2 \tag{12.2}$$

In Eq. (12.2), V is in volts, e in coulombs, m in kilograms and v in metres per second. Combining Eqs (12.1) and (12.2) we get

$$\frac{e}{m} = \frac{E^2}{2VB^2} \tag{12.3}$$

Thus, we can determine the charge-mass ratio of an electron, if we know E, V and B for the condition of zero deflection of the beam. Thomson's experiments (done somewhat differently from the method described here) gave a value of $e/m = 1.76 \times 10^{11}$ C kg^{-1}. In this experiment, only the ratio e/m could be measured.

Note: It is not possible to measure the absolute value of e, the electronic charge, by this method. Only the ratio e/m could be determined. The value of e was measured by an American physicist R.A. Millikan by his famous oil-drop experiment described in the next section. For his experiment, he was awarded the Nobel Physics prize in 1923. Knowing the value of e from this experiment and the value of e/m from Thomson's experiment, the mass (m) of the electron was known.

Example 12.1 Calculate the velocity of electrons accelerated by a potential of 10,000 V. The electronic charge $= 1.6 \times 10^{-19}$ C and mass $= 9.11 \times 10^{-31}$ kg.

Solution:

$$e = 1.6 \times 10^{-19}\ \text{C}$$

$$m = 9.11 \times 10^{-31}\ \text{kg}$$

$$V = 10{,}000\ \text{V}$$

Using Eq. (12.2), we have

$$v = \sqrt{\frac{2eV}{m}}$$

$$= 5.93 \times 10^{7}\ \text{ms}^{-1}$$

Example 12.2 A monoenergetic electron beam with electron speed of 5.20×10^6 ms^{-1} is subjected to a magnetic field of 1.30×10^{-4} T normal to the beam velocity. What is the radius of the circle traced by the beam, given e/m for electron equals 1.76×10^{11} C kg^{-1}.

Solution: The radius of the circle is given by [see Eq. (5.50)]

$$r = \frac{v}{Be/m} = \frac{5.20 \times 10^6}{1.30 \times 10^{-4} \times 1.76 \times 10^{-11}}$$

$$= 0.227 \text{ m} = 22.7 \text{ cm}$$

Example 12.3 The deflecting plates in a Thomson's setup are 5.0 cm long, and 1.5 cm apart. The plates are maintained at a potential difference of 240 V. Electrons accelerated to an energy of 2.0 keV enter from one edge of the plates midway in a direction parallel to the plates.

(a) What is the deflection at the other edge of the plates?

(b) At what distance from the undeflected position on the screen does the beam strike if the screen is 30 cm away from the other edge of the plates?

Solution:

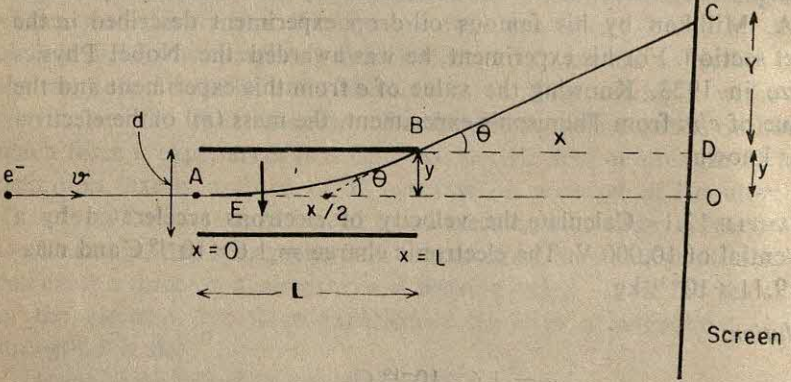


Fig. 12.8

(a) Let $x = L$ be the length of the plates and d their separation (Fig. 12.8). Let K be the kinetic energy of the electron beam entering the region between the plates at point A. The beam leaves the other edge of the plates at B suffering a deflection y. From Eq. (5.59), the deflection y is given by

$$y = \left(\frac{eE}{2mv^2}\right) \tag{i}$$

where v is the speed of electrons. The kinetic energy is

$$K = \frac{1}{2}mv^2$$

In terms of K, Eq. (i) reads

$$y = \frac{eEx^2}{4K} \tag{ii}$$

The electric field $E = V/d = 240\ \text{V}/1.5 \times 10^{-2}\ \text{m} = 1.6 \times 10^4\ \text{Vm}^{-1}$. Also $K = 2000\ \text{eV} = 2000 \times e$ joule and $x = L = 5.0 \times 10^{-2}$ m. Substituting these values in Eq. (ii) gives $y = 0.5$ cm.

(b) Let θ be the angle subtended with the axis of the plates by the tangent to the trajectory at point B. Then

$$\tan\theta = \left(\frac{dy}{dx}\right)_{\text{at } x=L}$$

Differentiating Eq. (ii) w.r.t. x we have

$$\tan\theta = \left(\frac{dy}{dx}\right)_{x=L} = \left(\frac{eEx}{2K}\right)_{x=L} = \frac{eEL}{2K}$$

or
$$\tan\theta = \frac{eEL^2}{4KL/2} = \frac{eEx^2}{4K}\bigg/\frac{x}{2} = \frac{y}{x/2}$$

This means that the tangent meets the axis at $x/2$, i.e. at $x = L/2$. When produced in the opposite direction, the tangent meets the screen at point C. It is clear from the diagram that

$$\tan\theta = \frac{y}{x/2} = \frac{Y}{X}$$

or
$$Y = \frac{2yX}{x} = \frac{2 \times 0.5 \times 30}{5.0} = 6.0\ \text{cm}$$

$\therefore$ Total deflection from centre O = OD + DC

$$= y + Y = 0.5 + 6.0 = 6.5\ \text{cm}$$

12.5 MILLIKAN'S OIL DROP EXPERIMENT FOR MEASURING ELECTRONIC CHARGE

Millikan's ingeneous experiment for the measurement of the fundamental charge is based on the study of the motion of charged oil drops under free fall due to gravity in a uniform electric field. His apparatus consisted of two horizontal circular metal plates A and B, about 22 cm in diameter and separated by a distance of about 16 mm, thus forming a parallel-plate capacitor (Fig. 12.9). The upper plate A had a small hole H in the middle. The plates were surrounded by a constant temperature bath D and the chamber C containing dry air. The plates were connected to a high-tension source so that potential differences of the order of 10,000 V could be established between the plates.

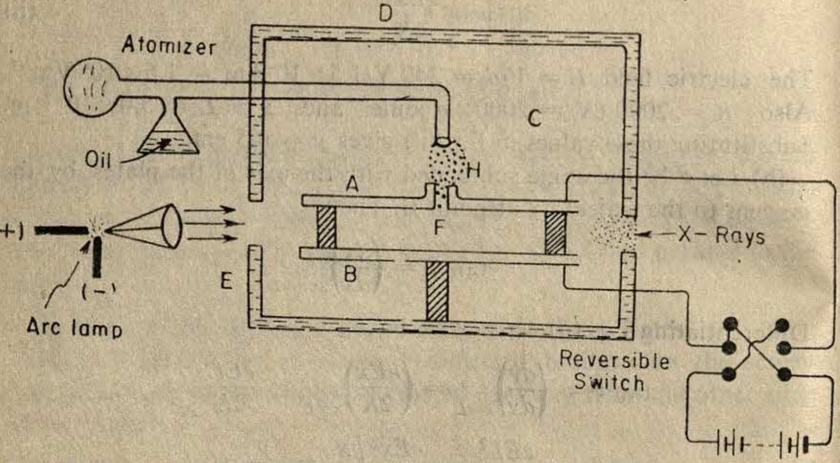


Fig. 12.9 *Millikan's oil drop experiment*

By means of an atomizer, a spray of fine droplets of a non-volatile liquid (such as glycerine) is produced near the hole H in the upper plate. The droplets move down slowly and occasionally a tiny drop makes its way through the hole H. The space F between the plates is strongly illuminated by an arc lamp. The oil droplets appear as specks of light against a dark background. The droplets are observed through a microscope whose eyepiece is provided with a micrometer scale (Fig. 12.10).

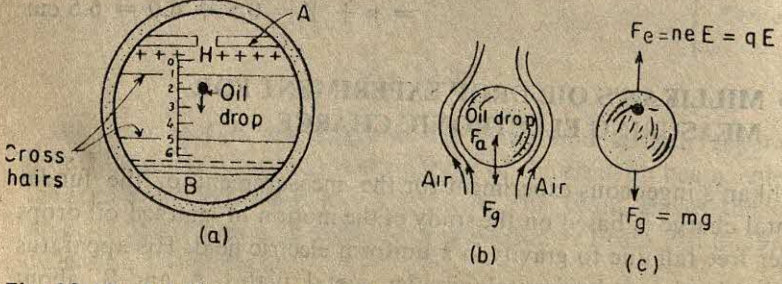


Fig. 12.10 *(a) Microscopic field of view*
(b) Drop in free fall
(c) Drop in electric field

The droplets get charged by friction in the spray. Additional charge can also be given by ionizing the air between the plates by means of x-rays.

An electric field was established between the plates with the upper plate connected to the positive terminal of a high-tension source. The magnitude of the field was adjusted so that a chosen oil drop moves slowly upward by the attraction of the upper plate. The uniform velocity v_e attained by the drop was determined by finding the time t_e taken by the drop to cover a known distance between the cross-hairs in the field of view of the microscope. The field was then switched off and the drop was allowed to fall freely under gravity. The uniform terminal velocity of the drop (v_g) was measured by finding the time (t_g) it took to travel a known distance between the cross hairs. By switching the field on and off, a particular oil drop could be kept in the field of view for a sufficiently long time.

You have learnt in Class XI that when a small sphere falls under gravity in a viscous fluid, it experiences an opposite force due to viscosity which increases with the increase in velocity of the sphere. Soon a stage is reached when the force due to viscosity becomes equal to the weight of the body and hence, the body moves with a uniform velocity called the terminal velocity.

Motion Under Gravity Let a be the radius of the oil drop and ρ the density of oil. The effective weight of oil in air is given by

$$mg = \frac{4\pi}{3}a^3(\rho - \sigma)g$$

where σ = density of air.

When the drop falls under gravity with a terminal velocity v_g, its weight balances with the force due to viscosity, and from Stokes' law we have

$$mg = \frac{4\pi}{3}a^3(\rho - \sigma)g = 6\pi\eta a v_g \tag{12.4}$$

where η is the coefficient of viscosity of air. Therefore, the radius of the drop is given by

$$a = \left[\frac{9}{2}\,\frac{\eta v_g}{(\rho - \sigma)g}\right]^{1/2} \tag{12.5}$$

Motion under Electric Field Let q be the magnitude of the negative charge on the drop and E the strength of the electric field. The upward force on the drop is qE. When the drop moves with a terminal velocity v_e, we have from Stokes' law

$$qE = mg + 6\pi\eta a v_e \tag{12.6}$$

because, in this case, the force due to viscosity acts downward opposite to the direction of motion of the drop. Equations (12.4) and (12.6) give

$$q = \frac{6\pi\eta a(v_g + v_e)}{E} \tag{12.7}$$

where the radius of the drop (a) is given by Eq. (12.5). Using this equation in Eq. (12.7) then gives

$$q = \frac{18\pi\eta(v_g + v_e)}{E}\left[\frac{\eta v_g}{2(\rho - \sigma)g}\right]^{1/2} \tag{12.8}$$

All quantities on the right hand side being known, the value of q can be calculated.

From his numerous observations of various droplets of different sizes having different charges Millikan discovered a remarkable fact. He found that each time the charge was an integral multiple (ne) of a certain minimum quantity e that is the fundamental charge or the charge on an electron. He found that the value of this elementary charge is $e = 1.6 \times 10^{-19}$ C.

Millikan also found that occasionally the velocity of an oil drop changed. He reasoned that this change was the result of the gain or loss of charge by the drop during its passage through the ionized air. Finding the charge in velocity he determined the charge gained or lost. The change in the charge was never found to be smaller than the smallest fundamental value e; it was always equal to an integral multiple of e. Thus Millikan's experiment established that *charge in nature is quantized.*

Knowing the value of e/m from Thomson's experiment and the value of e from Millikan's experiment, the mass (m) of the electron can be found to be

$$m = \frac{e}{e/m} = \frac{1.6 \times 10^{-19}\ \text{C}}{1.76 \times 10^{11}\ \text{C kg}^{-1}} = 9.11 \times 10^{-31}\ \text{kg}$$

EXAMPLE 12.4 In a Millikan's oil drop experiment, a charged oil drop of mass density 880 kg m^{-3} is held stationary between two parallel plates 6.0 mm apart held at a potential difference of 103 V. When the electric field is switched off, the drop is observed to fall a distance of 2.0 mm in 35.7 s.

(a) What is the radius of the drop?

(b) Estimate the charge of the drop. How many excess electrons does it carry? (The upper plate in the experiment is at a higher potential.)

(Viscosity of air $= 1.80\times10^{-5}$ Nsm^{-2}; $g = 9.81$ ms^{-2}; density of air $= 1.29$ kgm^{-3})

Solution:

$$\rho = 880 \text{ kgm}^{-3} \qquad E = \frac{103 \text{ V}}{6.0\times10^{-3}} = 1.72\times10^{4} \text{ Vm}^{-1}$$

$$v_g = \frac{2.0\times10^{3} \text{ m}}{35.7 \text{ s}} = 5.60\times10^{-5} \text{ ms}^{-1}$$

$$\eta = 1.80\times10^{-5} \text{ Ns m}^{-2}$$

$$g = 9.81 \text{ ms}^{-2} \quad \text{and} \quad \sigma = 1.29 \text{ kg m}^{-3}$$

(a) Substituting these values in Eq. (12.5), we find that the radius of the oil drop is

$$a = 7.26\times10^{-7} \text{ m}$$

(b) Since the charged drop is held stationary in the electric field, $v_e = 0$ so that Eq. (12.7) becomes

$$q = \frac{6\pi\eta a v_g}{E}$$

Substituting the values of the quantities involved gives

$$q = 8.05\times10^{-19} \text{ C}$$

Number of excess electrons is

$$n = \frac{q}{e} = \frac{8.05\times10^{-19}}{1.6\times10^{-19}} = 5$$

Thus the drop carries five excess electrons.

12.6 PHOTOELECTRIC EFFECT

We know that metals, when heated, emit electrons. Can electrons be released from metals by light? It has been observed that metals, when exposed to electromagnetic radiations such as x-rays, γ-rays, ultraviolet, visible, and infrared light, emit electrons. This phenomenon is called *photoelectric effect* and the emitted electrons are called *photoelectrons* because they are liberated by means of light. The photoelectric effect can be demonstrated by a very simple experiment as illustrated in Fig. 12.11. Light from a source is made to

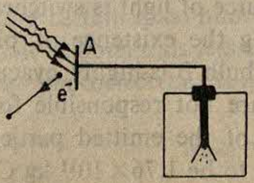


Fig. 12.11 *Demonstration of photoelectric effect*

strike an insulated metal plate A that is connected to an electroscope. Let us suppose that the electroscope is negatively charged.

When the light is turned on, its leaves are seen to collapse. However, if the electroscope were positively charged, its leaves would remain divergent. This can happen only if plate A acquires a positive charge when exposed to light, implying that electrons must have been ejected from it by the interaction of light with the matter of the plate. Further studies revealed that different substances emit photoelectrons only when exposed to different kinds of radiations. For example, alkali metals like potassium and cesium will eject photoelectrons when visible light falls on them. Heavier metals require shorter wavelengths of ultraviolet light for photoelectric emission. X rays will eject photoelectrons from the inner (K and L) shells of heavy elements. Cesium is sensitive even up to the infrared region. Ultraviolet light will eject photoelectrons from any metal.

Laws of Photoelectric Emission

The important features of photoelectric effect can be studied by a simple experimental arrangement sketched in Fig. 12.12. Two zinc plates P and Q, enclosed in an evacuated quartz bulb B, are connected to a battery and a microammeter as shown in the figure. When the plate P is connected to the negative terminal of the battery and ultraviolet light falls on it, the microammeter indicates a current in the circuit. But when the light is incident on the positive plate Q there is no flow of current. Hence, it is the negatively charged plate that loses its charge when ultraviolet light falls on it. The observed current in the circuit can be explained by the fact that the electrons emitted by the negatively charged plate P are attracted by the positive plate Q. The flow of electrons from P to Q constitutes a current. When the source of light is switched off, no current flows in the circuit establishing the existence of photoelectric effect. This effect occurs when the bulb B is highly evacuated, indicating that the air molecules in B are not responsible for the observed effect. The charge to mass ratio of the emitted particles is subsequently measured. The ratio is found to be 1.76×10^{11} kg C^{-1} in agreement with that for electrons. This confirms that the emitted particles are indeed electrons.

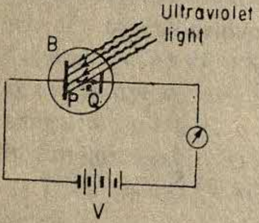


Fig. 12.12 *Experimental arrangement to study photoelectric effect*

Dependence of Photoelectric Current on Intensity of Incident Light

Further investigations by Lenard revealed that as long as the spectral quality of the light remains the same, the photoelectric current is directly proportional to the intensity of the incident light. If I is the photoelectric current, the number of photoelectrons emitted per second is I/e, where e is the electronic charge. Since e is a constant, it follows that the number of photoelectrons emitted per second is directly proportional to the intensity of the incident light.

Kinetic Energy of Photoelectrons

Richardson and Compton measured photoelectric current at various potentials V on Q relative to P, giving small positive and negative values to V but using light of the same spectral quality, i.e. wavelength. They repeated the experiment for various intensities of the incident light. Their results are depicted in Fig. 11.13. A study of these curves reveals that: (i) the photoelectric current reaches a saturation value above a certain value of V, the saturation current being proportional to the intensity of the incident light, and (ii) the photoelectric current decreases as negative voltage V is increased until it becomes zero at a negative voltage V_0. Below V_0 no current is observed no matter what the intensity of light is. V_0 is the same for the three curves which represent increasing intensity of the incident light.

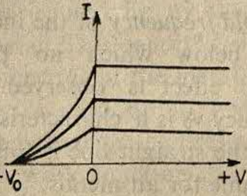


Fig. 11.13 *Photoelectric current at various voltages between the electrodes; the three curves are for increasing intensity of the incident light*

The cut-off potential V_0 depends on the nature of the emitter, i.e. the metal, and the frequency ν or wavelength λ of the incident light. For a given emitter illuminated by a monochromatic (same ν or λ) light, the photoelectrons are emitted with velocities ranging from zero up to a certain maximum value $v_{\max}$. Since there is no photoelectric current at potentials less than V_0, the maximum velocity $v_{\max}$ acquired by the photoelectrons is given by

$$K_{\max} = \frac{1}{2}mv_{\max}^2 = eV_0 \tag{12.9}$$

where $K_{\max}$ $(= \frac{1}{2}mv_{\max}^2)$ is the maximum kinetic energy. When V is between V_0 and zero, only those photoelectrons whose velocity satisfies $\frac{1}{2} mv^2 > eV$, can reach the anode. The electrons with smaller velocity than v are repelled back by the negative potential of the anode. When $V = 0$, the photoelectrons with velocity greater than zero can reach the anode.

Relationship Between the Cut-off Potential and Frequency of Light

The maximum energy of emission K_{max} $(= eV_0)$ was measured using lights of various frequencies ν. Figure 12.14 shows how eV_0 varies with ν. It was observed that for a given metal, there is a certain frequency ν_0, called the *threshold frequency* of the incident light, below which no photoelectric effect is observed. The frequency ν_0 is a characteristic of the photo-emitting metal. The slope of the straight line is found to be $= h$, the Planck's constant, and is the same for all metals.

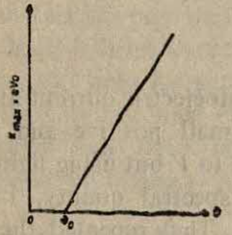


Fig. 12.14 *Threshold frequency*

To summarize, the laws of photoelectric emission are:

1. For a given emitter illuminated by light of a given frequency, the number of photoelectrons is directly proportional to the intensity of the incident light.

2. For every emitter there is a definite threshold frequency ν_0 below which no photoelectrons are emitted, no matter what the intensity of light is.

3. Above the threshold frequency, the maximum kinetic energy of the photoelectrons is proportional to the frequency of the incident light.

Einstein's Photoelectric Equation

The classical electromagnetic wave theory of light, which successfully explained interference, diffraction, and polarization of light, could not account for the observations related to photoelectric effect. In 1900 Planck postulated that light waves consist of tiny bundles of energy called *photons* or *quanta*. The energy of a light wave of frequency ν is given by $E = h\nu$, where h is the Planck's constant. Photon is simply a light wave of energy E.

Following Planck's idea, Einstein proposed a theory for photoelectric effect. According to him, when a photon of light falls on a metal, it is absorbed, resulting in the emission of a photoelectron. The kinetic energy $K = \frac{1}{2} mv^2$ of the emitted electron is given by

$$\frac{1}{2} mv^2 = h\nu - W_0 \tag{12.10}$$

This is the famous Einstein's photoelectric equation. The term $h\nu$ represents the total energy of the photon incident on the metal surface, as shown in Fig. 12.15. The photon penetrates a distance of about 10^{-8} m before it is completely absorbed. In disappearing, the photon imparts all its energy to a single electron. Part of this energy is used up by the electron in freeing itself from the atoms of the metal. This energy designated by W_0 in Eq. (12.10) is called the *work-function* of the metal and is a characteristic of it. The rest of the energy is used up in giving the electron kinetic energy $\frac{1}{2} mv^2$. The emitted electron acquires a velocity v.

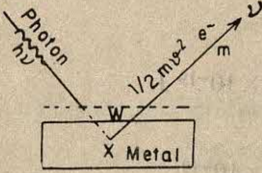


Fig. 12.15 *Explanation of emission of photoelectrons*

The work function W_0, i.e. the energy required to pull an electron away from the surface of the metal, is large for heavier elements like platinum whereas for other elements like alkali metals, W_0 is quite small. The minimum, or threshold, energy which a photon must have to free the electron from the surface of the metal should be equal to its work function. If the threshold frequency is ν_0 the threshold energy $= h\nu_0$. Thus

$$W_0 = h\nu_0$$

Einstein's photoelectric equation, therefore, becomes

$$\frac{1}{2} mv^2 = h(\nu - \nu_0) \tag{12.11}$$

It is evident that when $\nu < \nu_0$, no electron is emitted for any intensity of light. When $\nu > \nu_0$, the energy of the electron increases linearly with the frequency ν of light, which explains the experimental observations of Fig. 12.11. Since intensity of light is a measure of the number of photons and since each photon emits a photoelectron on absorption, the intensity of photoelectrons is proportional to the intensity of light, which explains the first law of photoelectric emission. Thus Einstein's theory accounts for the experimental observations on photoelectric effect.

EXAMPLE 12.5 A metal, whose work, function is 3.0 eV, is illuminated by light of wavelength 3×10^{-7} m. Calculate

(a) the threshold frequency
(b) the maximum energy of photoelectrons
(c) the stopping or retarding potential

Solution:

$$W_0 = h\nu_0$$
$$= 3.0\times 1.6\times 10^{-19}\ \text{J}$$

(a) Threshold frequency,

$$\nu_0 = \frac{W_0}{h} = \frac{3.0\times 1.6\times 10^{-19}}{6.63\times 10^{-34}}$$
$$= 0.72\times 10^{15}\ \text{Hz}$$

(b) Maximum energy

$$E_{\max} = \frac{1}{2}mv^2 = h(\nu - \nu_0)$$

Since $\lambda = 3\times 10^{-7}$ m,

$$\nu = \frac{c}{\lambda} = \frac{3\times 10^8}{3\times 10^{-7}} = 1.0\times 10^{15}\ \text{Hz}$$

$$\therefore \quad E_{\max} = h(\nu - \nu_0)$$
$$= 6.63\times 10^{-34}\times(1.0 - 0.72)\times 10^{15}$$
$$= 1.86\times 10^{-19}\ \text{J}$$
$$= 1.16\ \text{eV}$$

(c)
$$E_{\max} = eV_0$$
$$V_0 = \text{stopping potential}$$

$$\therefore \quad V_0 = \frac{E_{\max}}{e}$$
$$= \frac{1.86\times 10^{-19}}{1.6\times 10^{-19}} = 1.16\ \text{V}$$

EXAMPLE 12.6 (a) An x-ray tube produces a continuous spectrum of radiation with its short-wavelength end at 0.45 Å. What is the maximum energy of a photon in the radiation?

(b) From your answer to (a), guess what order of accelerating voltage (for electrons) is required in such a tube?

Solution:

(a)
$$\lambda_{\min} = 0.45 \times 10^{-10}\ \text{m}$$

$$E_{\max} = h\nu_{\max} = \frac{hc}{\lambda_{\min}}$$

$$= \frac{6.63 \times 10^{-34} \times 3.0 \times 10^{8}}{0.45 \times 10^{-10}}$$

$$= 4.42 \times 10^{-15}\ \text{J}$$

$$= 27.6 \times 10^{3}\ \text{eV} = 27.6\ \text{keV}$$

(b) The minimum accelerating voltage for electrons

$$= \frac{27.6\ \text{keV}}{e} = 27.6\ \text{kV}$$

i.e. of the order of 30 kV.

12.7 THE PHOTO-CELL

A photo-cell is an arrangement which converts light energy into electrical energy. It measures the intensity of light by measuring the photoelectric current when light falls on a photo-sensitive metal. It consists of a thin glass or quartz bulb which is highly evacuated and contains a parabolic metal surface C coated with photo-sensitive metal such as molybdenum, caesium, rubidium, etc. This serves as the cathode. At the centre of the curved surface is fixed a metal rod A which serves as the anode and also as the collector of photoelectrons (Fig. 12.16).

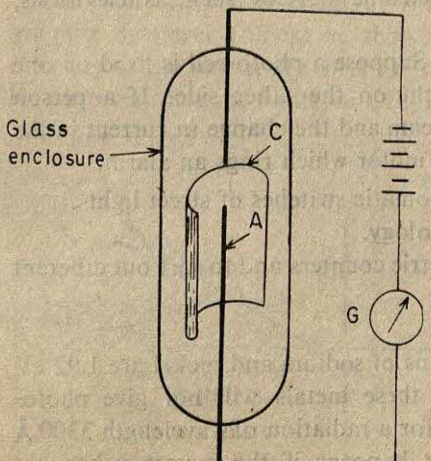


Fig. 12.16 *A photo-cell*

The anode is connected to the positive terminal of a battery and the cathode to the negative terminal. When light of frequency greater than the threshold frequency is made to fall on the photo-sensitive surface C (emitter), photoelectrons are emitted. These electrons are attracted to the positive anode A (collector). The current is measured by a sensitive galvanometer (or microammeter) included in the circuit. Since the number of electrons emitted depends upon the intensity of incident light, the photoelectric current registered by the galvanometer gives a measure of the intensity of light.

In the *vacuum type* cell described above the photoelectric current is very small, of the order of a microampere. To increase the current the cell is filled with a suitable inert gas, e.g. helmin, neon, etc. at a pressure of 0.1 mm to 1 mm of mercury. In the *gas filled type* cell, the current is enhanced due to the ionization of the gas by collisions of the gas molecules (atoms) with the ejected photoelectrons. In industry, gas-filled cells containing caesium oxide are very widely used, on account of their better response. They consist of a parabolic silver cathode, the surface of which is coated with a thin layer of caesium (about one molecule thick) in an atmosphere of oxygen and the anode is a metal rod mounted at the centre of the curved cathode surface.

Uses of Photo-cells

The basic function of a photo-cell is to convert a change in the intensity of illumination into a corresponding change in electric current. This change in electrical current is put to many uses described below.

1. to operate controls in electronic devices such as televisions, computers, etc.
2. to operate burglar alarms. Suppose a photo-cell is fixed on one side of a door and a source of light on the other side. If a person walks in, he interrupts the light beam and the change in current in the photo-cell may be used to start a motor which rings an alarm.
3. to switch on and off the automatic switches of street lights.
4. to record daylight in meteorology.
5. to count objects in photoelectric counters and to sort out different objects in photoelectric sorters.

EXAMPLE 12.6 The work functions of sodium and nickel are 1.92 eV and 5.0 eV respectively. Which of these metals will not give photoelectric current from a photo-cell for a radiation of wavelength 3300 Å from a source 1 m away? What happens if the source is brought nearer the cell at a distance of 50 cm?

Solution: Frequency of incident radiation is

$$\nu = \frac{c}{\lambda} = \frac{3.0 \times 10^8}{3300 \times 10^{-10}} = 9.1 \times 10^{14} \text{ Hz}$$

Threshold frequency of sodium is

$$\nu_0(\text{Na}) = \frac{1.92 \times 1.6 \times 10^{-19}}{6.63 \times 10^{-34}} = 4.6 \times 10^{14} \text{ Hz}$$

and of nickel is

$$\nu_0(\text{Ni}) = \frac{5.0 \times 1.6 \times 10^{-19}}{6.63 \times 10^{-34}} = 12 \times 10^{14} \text{ Hz}$$

Since ν is greater than $\nu_0(\text{Na})$ but less than $\nu_0(\text{Ni})$, nickel will not show photoelectric emission. When the source is brought closer to the cell, the photoelectric current from sodium increases due to an increase in intensity of the incident radiation but nickel will not give any photoelectric emission, howsoever high the intensity, since $\nu_0(\text{Ni})$ is greater than ν.

12.8 COMPTON SCATTERING

Einstein was awarded the Nobel Prize in Physics in the year 1921 for his quantum theory of the photoelectric effect. A confirmation of the concept of photon (i.e. particle-like behaviour of light) was provided in 1923 by A.H. Compton who was awarded the Nobel Prize for his work in 1927.

Compton allowed a beam of monochromatic x-rays to fall on a graphite scatterer (block) as shown in Fig. 12.17, and measured the intensity of the scattered x-rays for various angles of scattering (ϕ). He observed that the scattered x-rays had a longer wavelength λ' than

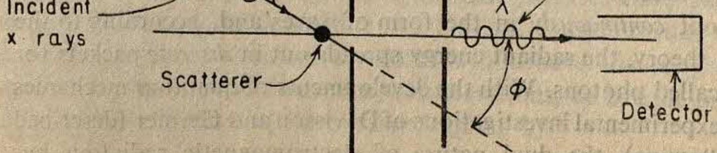


Fig. 12.17 *Schematics of Compton scattering experiment*

wavelength λ of the incident x-rays. The *Compton shift* $\Delta\lambda = (\lambda' - \lambda)$ varies with the angle ϕ at which the scattered x rays are observed.

The observed scattered wave of wavelength λ' cannot be understood if the incident x-rays are regarded as an electromagnetic wave. The incident wave of frequency ν (wavelength λ) would cause the electrons in the scattering block to oscillate at the same frequency. The oscillating electrons radiate electromagnetic waves again of the same frequency ν and hence the same wavelength λ. Thus, on the wave picture, the scattered x-rays should have the same wavelength as the incident x-rays.

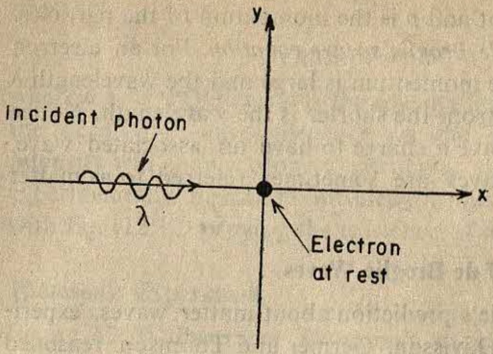
Compton explained his experimental results by postulating that the incident x-ray beam was not a wave but a group of photons of energy $E = h\nu$. As a result of collisions with the free electrons in the scattering block, the "recoil" photons emerging from the block form the scattered radiation. Since the incident photon transfers a part of its energy to the electron with which it collides, the scattered photon must have a lower energy E'. Hence it must have a lower frequency $\nu'(= E'/h)$ which implies a larger wavelength $\lambda'(= c/\nu')$ (see Fig. 12.18). This point of view accounts for the Compton shift $\Delta\lambda$ qualitatively. The quantitative expression for $\Delta\lambda$ is obtained by using the principles of conservation of energy and momentum with relativistic corrections. Here we simply quote the final result which reads:

$$\Delta\lambda = \frac{h}{m_0 c}(1 - \cos\phi)$$

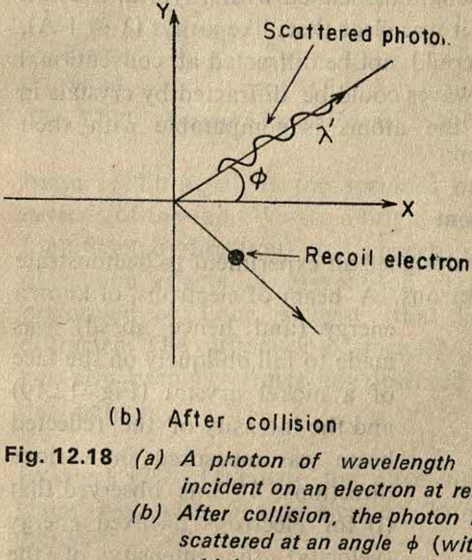
where ϕ is the angle of scattering of the x-ray photon and m_0 is the rest-mass of the electron.

12.9 WAVE NATURE OF MATTER

We have seen above that electromagnetic radiation has a dual nature. The experimental observations of interference, diffraction and polarization of light suggest that light is a wave. On the other hand, the experimental observations of photoelectric effect and Compton scattering experiments can be interpreted only in terms of the photon (quantum) theory. According to the wave theory, the radiant energy spreads out *continuously* in the form of waves and, according to the quantum theory, the radiant energy spreads out in *discrete* packets (or quanta) called photons. With the developments of quantum mechanics and the experimental investigations of Davisson and Germer (described in the following), the dual nature of electromagnetic radiation has been established.



(a) Before collision



(b) After collision

Fig. 12.18 *(a) A photon of wavelength λ incident on an electron at rest*
(b) After collision, the photon is scattered at an angle φ (with a higher wavelength λ') while the electron moves along another direction

In 1924, Louis de Broglie, a French theoretical physicist, derived an equation which predicted that all atomic particles have associated with them waves of a definite wavelength. Under certain circumstances, a beam of electrons or atoms will behave like a group of waves. On the basis of theoretical considerations, de Broglie predicted that the wavelength λ of these waves is given by

$$\lambda = \frac{h}{p} \qquad (12.12)$$

where h is Planck's constant and p is the momentum of the particles. This equation is known as *de Broglie's wave equation.* For an electron moving at a high speed, the momentum is large and the wavelength λ is small. The faster the electron, the shorter is the wavelength. Notice that the particle need not have a charge to have an associated wave. This is why de Broglie waves are sometimes referred to as matter waves.

Experimental Verification of de Broglie Waves

Immediately after de Broglie's prediction about matter waves, experimental physicists, notably Davisson, Germer and Thomson reasoned that a beam of electrons should show diffraction effects as ordinary light or x-rays. Since the wavelength associated with a typical electron beam is very short ($\lambda \sim 1$ Å), it was clear that, like x-rays ($\lambda \sim 1$ Å), matter waves, if they exist, could not be diffracted at conventional obstacles. Like x-rays, matter waves could be diffracted by crystals in which the spacings between the atoms is comparable with their wavelength.

Davisson and Germer Experiment

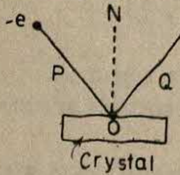


Fig. 12.19 *Diffraction of electrons*

In 1927, Davisson and Germer designed an experiment to demonstrate diffraction of a beam of electrons. A beam of electrons, of known energy (and hence speed) was made to fall obliquely on the face of a nickel crystal (Fig. 12.19) and the intensity of the reflected beam was measured in various directions. It was observed that (i) for electrons of fixed energy (velocity), the intensity of the reflected beam is maximum in the direction OQ and the angle subtended by OQ with the normal ON is the same as the angle subtended by the incident beam PO with the normal, and (ii) if the velocity of the incident electrons is changed gradually, the intensity of the reflected beam along OQ changes, becoming maximum and minimum depending on the speed of the electrons. In other words, in the same direction, the electrons of some speed are strongly reflected and those of a different speed are feebly reflected. To account for these observations, Davisson and Germer assumed that electrons behave like waves and that they were diffracted by the atoms of the crystal. The intensity of the reflected beam was found to be maximum for values of θ

given by

$$2d \sin \theta = n\lambda \tag{12.13}$$

where $\theta = \angle POQ$ and d is the spacing between atoms. Since the λ of the de Broglie waves depends on the velocity of the particle, the intensity will change along a particular direction with velocity. The experimentally measured wavelength was found to be in conformity with Eq. (12.12) proving the existence of de Broglie waves.

Thomson's Experiment

Figure 12.20 illustrates Thomson's experiment to demonstrate the existence of de Broglie waves. Following the experiment of Laue of the diffraction of x-rays, Thomson sent a beam of electrons on a platinum foil F and received the emergent electrons on a photographic plate P. The electron beam is diffracted at the spacings between the randomly oriented crystals of the foil. We see a bright spot at the centre of the plate (undiffracted direction) and concentric circles around it, indicating diffraction effects. A comparison of this experiment with the one performed by Laue suggests that like x-rays, electrons are also diffracted. The diffraction of uncharged particles, e.g. helium atoms, has also been experimentally observed. Thus, matter is endowed with dual nature.

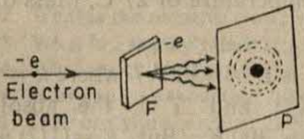


Fig. 12.20 *Thomson's experiment*

EXAMPLE 12.7 Calculate the wavelength of de Broglie waves associated with electrons accelerated through a potential difference of 200 V.

Solution: The energy gained by an electron

$$\begin{aligned} E &= eV \\ &= 1.6 \times 10^{-19} \times 200 \\ &= 3.2 \times 10^{-19} \text{ J} \end{aligned}$$

This energy is in the form of kinetic energy $\frac{1}{2} mv^2$

$$\therefore \qquad \frac{1}{2} mv^2 = eV$$

or

$$v = \sqrt{\frac{2eV}{m}}$$

$$= \sqrt{\frac{2 \times 3.2 \times 10^{-19}}{9.11 \times 10^{-31}}} = 8.4 \times 10^{6}\ \mathrm{ms^{-1}}$$

$$\therefore \qquad \lambda = \frac{h}{p} = \frac{h}{mv}$$

$$= \frac{6.63 \times 10^{-34}}{9.11 \times 10^{-3} \times 8.4 \times 10^{6}}$$

$$= 0.86 \times 10^{-10}\ \mathrm{m} = 0.86\ \text{Å}$$

EXAMPLE 12.8 Calculate the wavelength of de Broglie waves associated with thermal neutrons at room temperature of 27°C. Mass of a neutron $= 1.675 \times 10^{-27}$ kg.

Solution: Kinetic energy of neutrons is $E = (3/2)kT$ where k is the Boltzmann constant $= 1.38 \times 10^{-23}\ \mathrm{JK^{-1}}$ and T is the absolute temperature of neutrons $= 273 + 27 = 300$ K. But $E = (1/2)\,mv^2$. Therefore

$$\frac{1}{2} mv^2 = \frac{3}{2} kT$$

$$\text{or} \qquad p = mv = \sqrt{3mkT}$$

$$\therefore \qquad \lambda = \frac{h}{p} = \frac{h}{\sqrt{3mkT}}$$

Substituting the values of h, m, k and T gives

$$\lambda = 1.45 \times 10^{-10}\ \mathrm{m} = 1.45\ \text{Å}$$

SUMMARY

The phenomenon of the emission of electrons, when metals are exposed to light, is known as photoelectric effect. This effect can be explained only on the basis of the quantum theory, according to which light waves consist of tiny packets of energy called photons. Einstein gave the theory of the effect.

The dual nature of matter was predicted by de Broglie in 1924. According to him, a moving particle, under certain conditions, behaves like a wave. The experiments of Davisson, Germer, and Thomson verified the existence of de Broglie waves.

EXERCISES

A. Short-Answer Questions

1. What is the nature of cathode rays?
2. What is the magnitude of the fundamental charge?
3. What is charge to mass ratio of an electron?
4. What are photo-electrons?
5. What is de Broglie wavelength?

B. Long-Answer Questions

1. Describe the general appearance of the discharge in a tube when the pressure is gradually lowered.
2. What is the negative glow? What is the positive column?
3. What is Faraday's dark space and Crooke's dark space?
4. What are cathode rays?
5. How can you show that cathode rays travel in straight lines?
6. How can you show that cathode rays have momentum and energy?
7. How can it be shown that cathode rays are charged particles and that the charge is negative?
8. Describe how the charge of the electron can be experimentally determined.
9. Describe Thomson's experiment for the measurement of e/m of the electron.
10. What is photoelectric effect? How do photoelectrons differ from cathode rays?
11. What are the laws of photoelectric emission?
12. What is Planck's constant? Give its value with unit.
13. What are light quanta? What is a photon?
14. Give Einstein's explanation of photoelectric effect. Discuss the meaning of each term in his photoelectric equation.
15. What is the work function of a metal?
16. Describe an experiment to show that a moving electron has an associated wavelength. What determines its wavelength?

C. Multiple-Choice Questions

Choose the correct answer from the given alternatives.

1. Cathode rays of velocity 10^6 m/s describe an approximate circular path of radius 1 m in an electric field of 300 V/cm. If the velocity of cathode rays is doubled, the value of the electric field needed so that the rays describe the same circular path is

   (a) 1200 V/cm (b) 200 V/m
   (c) 600 V/cm (d) 2400 V/cm

2. An electron is not deflected on passing through a certain region, because

   (a) there is no magnetic field in that region
   (b) there is a magnetic field but velocity of the electron is parallel to the direction of magnetic field
   (c) the electron is a chargeless particle
   (d) none of the above

3. The frequency of the charge criculating at right angles to a uniform magnetic field does not depend upon

(a) the speed of the charge
(b) the mass of the charge
(c) the charge of the particle
(d) the magnetic field

4. The mass of a proton is 1847 times that of an electron. An electron and a proton are projected into a uniform electric field in a direction at right angles to the direction of the field with the same initial kinetic energy. Then

(a) the electron trajectory will be less curved than the proton trajectory.
(b) the proton trajectory will be less curved than the electron trajectory.
(c) both the trajectories will be equally curved
(d) none of the above.

5. Two electron beams having velocities in the ratio of 1 : 2 are subjected separately to identical magnetic fields. The ratio of deflections produced is

(a) 4 : 1
(b) 1 : 2
(c) 1 : 4
(d) 2 : 1

6. You are sitting in a room in which a uniform magnetic field is present in the vertically downward direction. When an electron is projected in horizontal direction, it will be moving in a circular path with constant speed

(a) clockwise in the vertical plane
(b) clockwise in the horizontal plane
(c) anti-clockwise in the horizontal plane
(d) anti-clockwise in the vertical plane

7. The work function for aluminium surface is 4.2 eV. The cut-off wavelength for the photoelectric effect for the surface is

(a) 2955 Å
(b) 3100 Å
(c) 4200 Å
(d) 1500 Å

8. The work function for an aluminium surface is 4.2 eV and that for a sodium surface is 2.0 eV. The two metals were illuminated with appropriate radiations so as to cause photo emission. Then

(a) both aluminium and sodium will have same threshold frequency
(b) the threshold frequency of aluminium will be more than that of sodium
(c) the threshold frequency of aluminium will be less than that of sodium
(d) none of the above

9. A caesium photo-cell with a steady potential difference of 60 V across it, is illuminated by a small bright light placed 50 cm away. When the same light is placed one meter away the electrons crossing the photo cell

(a) each carry one quarter of their previous energy
(b) each carry one quarter of their previous momentum
(c) are half as numerous
(d) are one quarter as numerous

10. An image of the sun is formed by a lens of focal length 30 cm on the metal surface of a photoelectric cell and it produces a current *I*. The lens formni-

the image is then replaced by another lens of the same diameter but of focal length 15 cm. The photoelectric current in this case will be

(a) $I/2$ (b) $2I$
(c) I (d) $4I$

D. Numerical Problems

1. A narrow beam of electrons, accelerated by a potential difference of 250 V, passes between two parallel metal plates 0.030 m long and 0.005 m apart in an evacuated tube, and then strikes a fluorescent screen 0.175 m from the centre of the plates. A potential difference of 10 V is established between the plates. What is the deflection produced on the screen?
2. In a cathode ray tube a potential difference of 3000 V is maintained between the electrodes 2 cm apart. A magnetic field of 2.5×10^{-4} Wb/m^2 at right angles to the electric field gives no deflection of the electron beam which received an initial acceleration by a potential difference of 10000 V. Calculate the value of e/m.
3. A charged oil drop is held stationary between two horizontal plates of a capacitor when a field of 576 kV/m is applied between them. When the field is removed the drop falls freely with a steady velocity of 0.120×10^{-3} m/s. Calculate the radius and charge on the drop. Given that $\eta = 0.109\times10^{-4}$ kg m^{-1} s^{-1} and the difference between the densities of oil and air may be taken as 800 kg m^{-3}.
4. The most energetic electrons emitted from a surface by 3500 Å photons are found to be bent in a 0.18 m circle by a magnetic field of 1.5×10^{-5} Wb/m^2. Calculate the work function for the material.
5. Photoelectrons, of charge e and constant mass m are liberated in a vacuum and with negligible velocities from the negative plate of a parallel plate capacitor. A magnetic field **B** is applied parallel to the plates. The separation between the plates is x. Show that a potential difference

$$V = \left(\frac{eB^2}{2m}\right)x^2$$

when applied across the capacitor will enable the electrons to just reach the positive plate.
6. Light of wavelength 4500 Å is incident on two photoelectric tubes. The emitter in the first tube has a threshold wavelength of 6000 Å and the emitter in the second tube has a work function twice as large as in the first tube. Find the stopping potential in each of the tubes.
7. The kinetic energies of photoelectrons range from zero to 4.0×10^{-19} J, when light of wavelength 3000 Å falls on a surface. What is the stopping potential for this light?
8. Compare the energies of photons of wavelength 10^{-10} m and 6×10^{-7} m.
9. No photoelectrons are emitted from a metal if the wavelength of light exceeds 6000 Å. Calculate the work function of the metal.
10. Calculate wavelength of a photon of energy 10^{-19} J.
11. A metal of work function 5 eV is illuminated by monochromatic light of wavelength 1.2×10^{-7} m. Calculate

(a) the threshold frequency
(b) the threshold wavelength
(c) the maximum energy of photoelectrons
(d) the cut-off potential

12. The maximum energy of photoelectrons is 10×10^{-19} J when a photo-emitting metal is illuminated by a monochromatic light of wavelength 1.0×10^{-7} m. Calculate the work function of the metal.
13. A phototube has a cut-off potential of 1.0 V. Calculate the maximum velocity of the photoelectrons. What is the work function of the metal if it is illuminated by monochromatic light of wavelength 6×10^{-7} m?

13

ATOMS, NUCLEI AND MOLECULES

13.1 INTRODUCTION

J.J. Thomson's experiments with electric and magnetic fields revealed that the atom must consist of negatively charged particles (electrons) which break away from the cathode. Since the atom as a whole is electrically neutral, those atoms of the cathode that have lost electrons must have a residual positive charge. The total positive and negative charges in an atom must be exactly equal. How are these charges distributed in the atom? In an attempt to answer this question, J.J. Thomson proposed a model for the internal structure of the atom. According to him an atom consisted of a positively charged substance having electrons that are, more or less, uniformly distributed over the body of the substance like seeds in a watermelon (Fig. 13.1). This model was applied to predict the frequencies emitted by an incandescent solid (line spectrum). The theoretical predictions did not agree with experimentally observed frequencies, and it became clear that Thomson's picture of the atom was wrong.

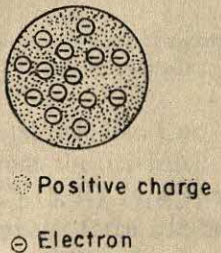


Fig. 13.1 *Thomson's model of the atom*

13.2 RUTHERFORD'S MODEL OF ATOM

The first correct description of the distribution of positive and negative charges within the atom was given by Ernest Rutherford (1871–1937). His atomic model was based on the epoch-making experiments of the scattering of α-particles from thin metal foils, performed by his students, Geiger and Marsden. An α-particle is a helium ion, i.e. a helium atom from which both its electrons are removed. It has a charge $2e$ and a mass that is nearly four times the mass of a proton. The mass of a proton is 1.67×10^{-27} kg. Fast moving α-particles are emitted (among other particles) in radioactive disintegrations. The basic idea of Rutherford's experiment is very simple (Fig. 13.2). S is a speck of a radioactive material, emitting energetic α-particles, that are collimated to a fine beam by a slit A. The beam is made to fall on a thin metal foil B($\sim 10^{-4}$ cm). The scattered (or deflected) particles are received on a fluorescent screen C. The tiny flashes of light, or scintillations, that are produced when an α-particle hits the screen are viewed with a microscope. The whole apparatus is enclosed in an evacuated chamber to prevent scattering of α-particles by gas molecules.

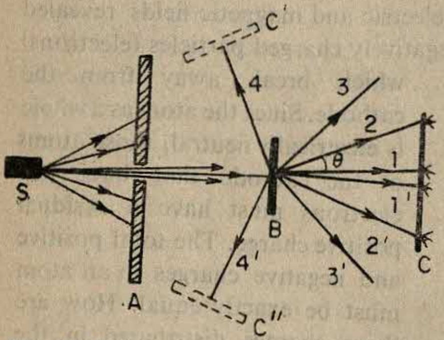


Fig. 13.2 *Scattering of α-particles by a thin metal foil*

Rutherford observed that most of the α-particles passed through the foil without any appreciable deflection. In Fig. 13.2 such particles are labelled 1, 1′. In other words, the foil behaved as if it were a wide mesh screen through which the particles pass. Now, a 10^{-4} cm thickness of the metal foil has several thousand atoms and such a compact pile of atoms cannot have wide gaps through which the α-particles can shoot. Rutherford reasoned that the atoms of the metal themselves must have gaps in them. Of the particles that were deflected in passing through the foil, many particles (such as 2, 2′, 3, 3′) suffered fairly large deflections θ. Some like 4, 4′ were deflected through more than 90°. Some particles even rebounded from the foil, suffering a deflection of nearly 180°. The screen placed at C′ or C″ detected such deflections. Thomson's model could predict only small deflections. Such large deflections require strong forces to be acting on the α-particles. Rutherford argued that this would be possible if all the positive charge and the mass of the atom were concentrated in a very small central region which he called the nucleus of the atom, as shown in Fig. 13.3.

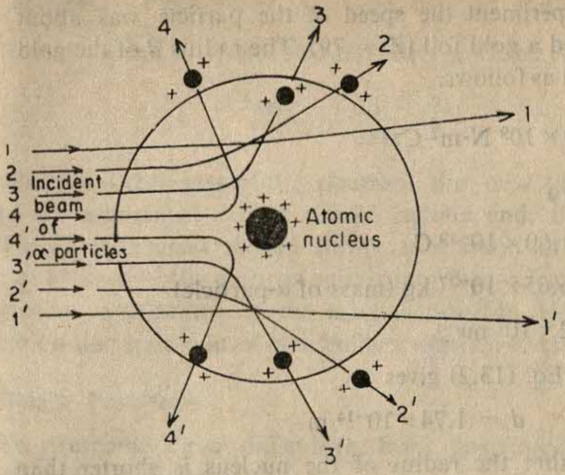


Fig. 13.3 *Deflection of α-particles by a small, massive positive nucleus*

An α-particle like 1 or 1′, penetrating the atom near its edge (i.e. away from the nucleus), would suffer only a small deflection. Those like (2, 2′), (3, 3′), (4, 4′) passing closer to the nucleus would be repelled with a greater force and hence deflected through a larger angle. The particle travelling head-on towards the nucleus is slowed down by it until it comes to rest at a distance *d* from it, and is then repelled back. The distance *d* of the closest approach of the fast α-particle, gives us a rough estimate of the radius of the nucleus. To compute *d* let us consider an α-particle of energy $E = \frac{1}{2}mv^2$, where v is the velocity of the particle. Due to Coulomb repulsion from the nucleus, the particle is slowed, thus losing its kinetic energy, which must change into its electrical potential energy. When the particle has reached the distance *d* of the closest approach, it is momentarily at rest. At this position the entire kinetic energy ($\frac{1}{2}mv^2$) has been converted into potential energy which is due to Coulomb repulsion of charge q_1 of the particle by the charge q_2 of the nucleus. Therefore, we have

$$\frac{1}{2}\,mv^2 = \frac{1}{4\pi\epsilon_0}\cdot\frac{q_1 q_2}{d} \qquad (13.1)$$

where $1/4\pi\epsilon_0$ is a constant and its value $= 9\times10^9$ N-m^2 C^{-2}. The charge q_2 of the nucleus of the atom of atomic number $Z = Ze$ and the charge q_1 of the particle is $2e$. Therefore,

$$d = \frac{1}{4\pi\epsilon_0}\cdot\frac{4Ze^2}{mv^2} \qquad (13.2)$$

In Rutherford's experiment the speed of the particle was about 2.0×10^{7} ms^{-1}. He used a gold foil ($Z = 79$). The radius d of the gold nucleus was computed as follows.

$$\frac{1}{4\pi\epsilon_0} = 9\times 10^{9}\ \text{N-m}^2\ \text{C}^{-2}$$

$$Z = 79$$

$$e = 1.60\times 10^{-19}\ \text{C}$$

$$m = 6.65\times 10^{-27}\ \text{kg (mass of } \alpha\text{-particle)}$$

$$v = 2\times 10^{7}\ \text{ms}^{-1}$$

Using these values in Eq. (13.2) gives

$$d = 1.74\times 10^{-14}\ \text{m}$$

Thus, he concluded that the radius of the nucleus is shorter than 1.7×10^{-14} m. Because the mass of an electron is only about 1/7000 of the mass of an α-particle, the effect of the presence of electrons in the atom; on the deflection of α-particles is negligible,

Rutherford's experiments suggest that the mass of the atom is not distributed uniformly throughout the atom. Most of the mass is located in a very small volume of radius $\sim 10^{-14}$ m, called the nucleus. The electrons are supposed to reside outside the nucleus.

13.3 ENERGY QUANTIZATION: BOHR'S MODEL OF HYDROGEN ATOM

In 1913, Niels Bohr proposed a theory of the hydrogen atom which not only gave a satisfactory explanation of the spectral lines of hydrogen but also a model for the structure of all other atoms. His starting point was the nuclear model of the atom, suggested by Rutherford. He assumed that a hydrogen atom ($Z = 1$), the simplest of all atoms, consists of a nucleus with one positive charge, $+e$, and a single electron of charge, $-e$, revolving around it in a circular orbit of radius r (Fig. 13.4).

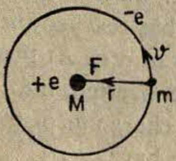


Fig. 13.4 *Bohr's model of hydrogen atom*

If the electron was at rest it would fall into the nucleus due to electrostatic attraction between the two oppositely charged particles. To keep the electron in a stable orbit, Bohr assumed that the inward centripetal force mv^2/r must be due to the inward electrostatic force $\frac{1}{4\pi\epsilon_0}\cdot\frac{e^2}{r^2}$. For a stable orbit. they must be equal, i.e.

$$\frac{mv^2}{r} = \frac{1}{4\pi\epsilon_0}\cdot\frac{e^2}{r^2} \tag{13.3}$$

where m is the mass of the electron. But now two difficulties arose: (i) an accelerated charge should radiate and, therefore, lose energy. The electron orbit should shrink as a result of loss of energy (see Eq. 13.3) until the electron spirals into the nucleus. (ii) As in planetary motion, any radius of the orbit is possible. How can such an atom emit a line spectrum of well-defined spectral lines (see Chapter 6)?

Bohr's Postulates

To overcome these difficulties, Bohr postulated new revolutionary ideas. His postulates are:

1. The conventional classical theory does not hold in the case of such tiny particles.
2. The electron in a stable orbit does not radiate energy as in the classical theory.
3. The electron cannot move in all the possible orbits allowed by the classical theory but only in certain definite and discrete orbits in accordance with the prescribed rules called quantum conditions.
4. The atom can absorb or radiate energy if the electron jumps from a lower to a higher orbit or falls from a higher to a lower orbit, respectively.

Bohr's third postulate is of far-reaching consequences. The existence of definite orbits implies discretization (or quantization) of energy of the hydrogen atom. Thus he laid the foundation of quantum physics.

The concept of quantization of physical quantities is not new; it also exists in classical physics. In Class XI you have learnt of the propagation of transverse waves on a string. You have seen that a wave of any wave-length (and frequency) will travel on a long string. But what happens if the string is not very long and if its ends are fixed as in a sonometer or a violin? Two important changes occur due to the boundary conditions: (a) the displacement of the string is now represented by *standing waves* rather than *travelling waves* and (b) only certain *discrete* wavelengths (or frequencies) can now exist. We have seen in Class XI that the allowed wavelengths are given by

$$\lambda_n = \frac{2L}{n} \qquad n = 1, 2, 3, \ldots \tag{13.4}$$

where L is the length of the string. This discretization or quantization of the wavelength is a direct consequence of putting boundaries or limitations on the wave. Sound waves in pipes exhibit a similar behaviour.

We now postulate that, just like a vibrating string, for a free particle of mass m confined to move on line l, the linear momentum p of the particle can have only those values that correspond to its de Broglie wavelength λ having only discrete values λ_n [Eq. (13.4)]. This condition gives

$$\lambda_n \left(= \frac{h}{p}\right) = \frac{2l}{n}$$

or

$$p = \frac{nh}{2l} \qquad n = 1, 2, 3, \ldots \tag{13.5}$$

where h is the Planck's constant. Equation (13.5) implies that the linear momentum of a particle confined to move on a line is quantized. In terms of momentum p, the energy of the particle is given by

$$E = \frac{1}{2}mv^2 = \frac{1}{2}\frac{m^2v^2}{m} = \frac{p^2}{2m}$$

Substituting for p from Eq. (13.5) we have

$$E_n = \frac{n^2h^2}{8ml^2} \qquad n = 1, 2, 3, \ldots \tag{13.6}$$

where E_n is the value of E for a particular value of $n = 1, 2, 3, \ldots$ i.e.

$$E_1 = \frac{h^2}{8ml^2}, \qquad E_2 = \frac{4h^2}{8ml^2},$$

$$E_3 = \frac{9h^2}{8ml^2}, \qquad \ldots$$

Thus, a free particle confined to a line can have only discrete values of energy corresponding to integral values $n = 1, 2, 3, \ldots$ These values of n label the stationary energy states in ascending order of energy.

Bohr's Quantization Condition

Let us now consider the motion of an electron in a circular orbit of radius r round the nucleus of the atom. The motion of the electron is confined to a circular path of length $2\pi r$. By analogy with the linear motion of a particle confined to a line, Bohr's reasoned that, in circular motion, the angular momentum (and hence energy) of the electron will become quantized, i.e. they can have only certain discrete values.

De Broglie succeeded in obtaining the quantization condition of the postulate (3) of Bohr by applying the appropriate boundary conditions to matter waves in the hydrogen atom. Figure 13.5 shows how an electron wave can be adjusted in wavelength to fit an integral number of times around the circumference of a Bohr orbit of radius r. The de Broglie wavelength $\lambda(= h/p)$ has been chosen such that the circumference $2\pi r$ of the orbit of radius r contains an integral number n of matter waves, or

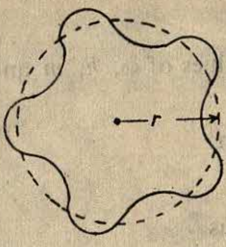


Fig. 13.5 *Instantaneous "snapshot" of a standing matter wave associated with an electron orbit of radius r*

$$\frac{2\pi r}{\lambda} = n$$

But $\lambda = h/p$, where h is the Planck's constant and p is the linear momentum (mass × velocity). Thus

$$\frac{2\pi r}{h/p} = n$$

or

$$L = pr = \frac{nh}{2\pi} \qquad n = 1, 2, 3, \ldots \tag{13.7}$$

where L is the angular momentum of the electron. Equation (13.7) is called Bohr's quantization condition for angular momentum. Bohr introduced this condition as an *ad hoc* postulate. It was a revolutionary concept as it violates the known ideas of classical physics. It restricts the stationary states of the electron to those circular orbits in which its angular momentum is an integral multiple of $h/2\pi$.

Bohr's next postulate (4) was equally revolutionary as it violates classical electrodynamics. He postulated that the accelerated motion of an electron in an orbit is not accompanied by the emission of electromagnetic radiation. By this postulate, Bohr provided stability to the hydrogen atom.

Radii of Bohr Orbits of Hydrogen

The radius of the electron orbit is governed not only by Eq. (13.3) but also by Bohr's quantization condition (13.7) which reads

$$pr = mvr = \frac{nh}{2\pi} \tag{13.8}$$

Eliminating v from Eqs (13.3) and (13.8) we have

$$r = \frac{\epsilon_0 n^2 h^2}{\pi m e^2} \tag{13.9}$$

Substituting in this equation, the known values of ϵ_0, h, m and e we obtain the radius r_n of the nth orbit:

$$\begin{aligned} r_n &= (0.529 \times 10^{-10}) n^2 \text{ m} \\ &= 0.529 n^2 \text{ Å} \end{aligned} \tag{13.10}$$

The innermost orbit with $n = 1$ has a radius

$$r_1 = 0.529 \text{ Å}$$

The second orbit ($n = 2$) is four times larger and the third ($n = 3$) nine times larger, and so on.

Using Eqs (13.8) and (13.9), the speed of the electron in the n-th orbit is given by

$$v_n = \frac{e^2}{2\epsilon_0 n h} = \frac{2.19 \times 10^6}{n} \text{ ms}^{-1} \tag{13.11}$$

Since the circumference of the nth circular orbit is $2\pi r_n$, the frequency with which the electron goes around the orbit is given by

$$\nu_r = \frac{v_n}{2\pi r_n} \tag{13.12}$$

It should be noted that a hydrogen atom has only one electron and that it can occupy only one orbit at any one time. The radius of the lowest orbit ($n = 1$) is 0.53 Å. The permitted orbits are given by the positive integral values of n, which is called the *principal quantum number*.

Energy of the Hydrogen Atom in nth State

Let us now compute the energy of the hydrogen system for a given n. The energy consists of the electron's kinetic energy and the electrostatic potential energy of the two charges. The kinetic energy of the electron in the nth orbit is given by [from Eq. 13.3]

$$\text{KE} = \frac{1}{2} m v_n^2 = \frac{1}{8\pi\epsilon_0} \frac{e^2}{r_n}$$

and its electrostatic energy is $\text{PE} = -e^2/(4\pi\epsilon_0 r_n)$ so that the total energy is given by (using Eq. 13.9)

$$E_n = \text{KE} + \text{PE} = -\frac{1}{8\pi\epsilon_0} \cdot \frac{e^2}{r_n} = -\frac{m e^4}{8\epsilon_0^2 n^2 h^2} \tag{13.13}$$

Thus, each Bohr Orbit has a definite energy E_n. Each orbit is regarded as a possible state of the system. The state with $n = 1$ has the lowest possible (most negative) energy.

Bohr's postulate (4) states that system can change from one energy state to another, the change in energy being compensated for by the emission or absorption of energy. Suppose the atom changes from a state with $n = n_1$ to a state with $n = n_2$, n_2 being a *smaller* number than n_1. The decrease in energy is $= E_{n1} - E_{n2}$. Bohr assumed that this energy difference is emitted as light of frequency ν, governed by (see Fig. 13.6),

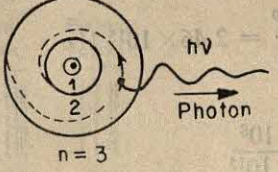


Fig. 13.6 *Schematic diagram of radiation of light from Bohr's hydrogen atom*

$$h\nu = E_{n1} - E_{n2} \tag{13.14}$$

Using Eq. (13.13), the frequency of emission, for any n_1 and n_2 is given by

$$\nu = \frac{E_{n1} - E_{n2}}{h} = \frac{me^4}{8\epsilon_0^2 h^3}\left(\frac{1}{n_2^2} - \frac{1}{n_1^2}\right) \tag{13.15}$$

Notice that the frequency ν of the light emitted is not the same as the frequency ν_r [Eq. 13.12)] with which the electron revolves in an orbit.

From Eq. (13.15), we have

$$\nu h = \frac{me^4}{8\epsilon_0^2 h^2}\left(\frac{1}{n_2^2} - \frac{1}{n_1^2}\right)$$

This is the *Rydberg formula* for the spectrum of the hydrogen atom.

The Origin of Spectrum

If the electron is in the smallest possible orbit $n = 1$, the atom is said to be in its *normal* or *ground* state. It will remain in this state for ever and will not emit light unless it is raised or "excited" to a higher state by the addition of energy from outside. The atom is then said to be in an *excited state*. The first excited state is $n = 2$. The excitation energy required to lift the electron from the state $n = 1$ to the state $n = 2$, can be computed from the equation,

$$E_{n_1} - E_{n_2} = \frac{me^4}{8\epsilon_0^2 h^2}\left(\frac{1}{n_2^2} - \frac{1}{n_1^2}\right) \tag{13.17}$$

by setting $n_2 = 1$ and $n_1 = 2$.

$$E_1 - E_2 = \frac{(2\pi)^2 me^4}{(4\pi\epsilon_0)^2 h^2}\cdot\left(-\frac{1}{4} + 1\right)$$

where $1/4\pi\epsilon_0 = 9 \times 10^9\ \text{Nm}^2\,\text{C}^{-2}$, $m = 9.11 \times 10^{-31}$ kg, $e = 1.6 \times 10^{-19}$ C and $h = 6.63 \times 10^{-34}$ Js. Substituting these values, we get

$$E_2 - E_1 = 16.29 \times 10^{-19}\ \text{J} = 10.21\ \text{eV}$$

From this excited state ($n = 2$), the electron can go back to the ground state ($n = 1$) by emitting a radiation of frequency ν given by

$$\nu = \frac{E_2 - E_1}{h} = \frac{16.29 \times 10^{-19}}{6.63 \times 10^{-34}} = 2.46 \times 10^{15}\ \text{Hz}$$

Wavelength of emitted radiation is

$$\lambda = \frac{c}{\nu} = \frac{3.0 \times 10^8}{2.46 \times 10^{15}}$$
$$= 1.22 \times 10^{-7}\ \text{m} = 1220\ \text{Å}$$

This wavelength lies in the ultraviolet region. The emitted radiation is seen as a line in the spectrum of hydrogen. The transitions from higher to lower orbits will give rise to other lines of the spectrum. To get a hydrogen ion, the electron must be completely removed from the atom, i.e. the electron must be lifted from $n_1 = 1$ state to $n_2 = \infty$ state. From Eq. (13.17) the energy required to ionize a hydrogen atom is found to be 13.6 eV, which is called the *ionization energy*.

The origin of the various possible lines of the spectrum of hydrogen atom can be shown graphically by marking by energies E_n of the various possible orbits. This is shown in Fig. 13.7 where the energy scale runs vertically and is graduated in electron volts, with zero energy at the top. The transitions from one state to another are indicated by arrows. The transitions from orbits higher than $n = 5$ have been omitted, for simplicity.

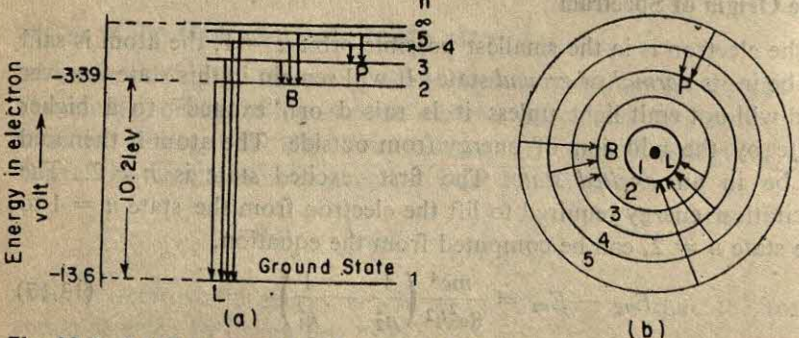


Fig. 13.7 (*a*) *The energy level diagram for the hydrogen atom; B—Balmer series, L—Lyman series, P—Paschen series*
(*b*) *Graphical representation of spectral series of hydrogen atom using circular electron orbits (diagram not to scale)*

The Spectrum of Hydrogen Atom

The spectrum of atomic hydrogen was observed much before Bohr proposed a theory for it. Figure 13.8 shows the line spectrum of hydrogen. The lines fall into three distinct groups or series, each named after its discoverer.

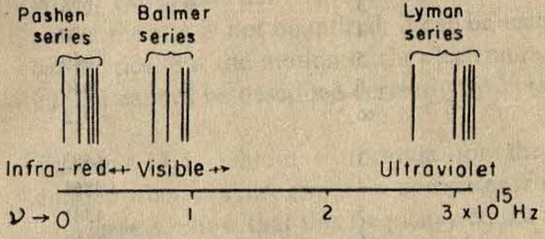


Fig. 13.8 *The spectrum of hydrogen atom*

The Balmer series lies in the visible part of the spectrum, the Paschen series in the infrared, and the Lyman series in the ultraviolet. The observed frequency of these lines matched exactly with that computed from Eq. (13.15). This indeed is a great success of Bohr's theory. The greatness of his revolutionary theory, that violated the established laws of the classical theory, lies in the fact that apart from successfully explaining the known series of hydrogen spectrum, the theory predicted new series which were later actually discovered by Brackett and Pfund.

Table 13.1 shows the wavelengths of some selected lines of the Lyman, Balmer and Paschen series of the hydrogen spectrum.

EXAMPLE 13.1 An electron is confined by electrical forces to move in a line between two walls separated by 5.0×10^{-10} m. Find the quantized energy values for the three lowest stationary states.

Solution: The energy of the *n*th stationary state is [see Eq. (13.6)]

$$E_n = \frac{n^2h^2}{8ml^2}$$

where $m = 9.11 \times 10^{-31}$ kg, $h = 6.63 \times 10^{-34}$ Js and $l = 5.0 \times 10^{-10}$ m.

Lowest energy state ($n = 1$)

$$E_1 = \frac{h^2}{8ml^2} = \frac{(6.63 \times 10^{-34})^2}{8 \times 9.11 \times 10^{-31} \times (5.0 \times 10^{-10})^2}$$

$$= 2.41 \times 10^{-19} \text{ J} = 1.5 \text{ eV}$$

The energies of the next two states ($n = 2$ and $n = 3$) are $(2)^2 \times 1.5$ eV $= 6.0$ eV and $(3)^2 \times 1.5$ eV $= 13.5$ eV respectively.

Table 13.1 The hydrogen spectrum (some selected lines)

Name of series	Quantum number		Wavelength
	i (lower state)	j (upper state)	10^{-10} m (or Å)
Layman	1	2	1216
	1	3	1026
	1	4	970
	1	5	949
	1	6	940
	1	∞	912
Balmer	2	3	6583
	2	4	4861
	2	5	4341
	2	6	4102
	2	7	3970
	2	∞	3650
Paschen	3	4	18751
	3	5	12818
	3	6	10938
	3	7	10050
	3	8	9546
	3	∞	8220

EXAMPLE 13.2 A dust particle of mass 1.0 μg is confined to move in a line between two walls separated by 0.1 mm. The speed of the particle is 10^{-6} ms^{-1}. Does this system belong to the realm of quantum physics? Give the reason for your answer.

Solution:

$$m = 1.0\ \mu\text{g} = 1.0 \times 10^{-6}\ \text{g} = 1.0 \times 10^{-9}\ \text{kg}$$

$$l = 0.1\ \text{mm} = 1.0 \times 10^{-4}\ \text{m}$$

$$v = 10^{-6}\ \text{ms}^{-1}$$

Let us find the quantum number n that describes the motion of the particle. The energy in the nth stationary state is

$$E_n = \frac{1}{2}mv^2 = \frac{1}{2} \times 10^{-9} \times (10^{-6})^2 = 5 \times 10^{-22}\ \text{J}$$

But

$$E_n = \frac{n^2h^2}{8ml^2}$$

$$\therefore \quad n = \sqrt{8mE}\ _n\frac{l}{h}$$

Substituting the values of m, E_n, l and h, we get

$$n = 3 \times 10^{14}$$

The value of n is so large that no experiment will be able to distinguish between n and $(n + 1)$ energy states. Hence this problem does not belong to the realm of the quantum physics; classical physics works extremely well. It is for this reason that the motion of planets round the sun is not quantized; it can be accurately described by classical physics. But the motion of the electron in Example 13.1 is quantized and cannot be described correctly by classical physics.

EXAMPLE 13.3 Obtain a formula for the frequency of radiation emitted when a hydrogen atom de-excites from level n to level $(n-1)$. For large n, show that this frequency equals the classical frequency of revolution of the electron in the orbit.

Solution: From Eq. (13.15) we have for $n_2 = n - 1$ and $n_1 = n$,

$$\nu = \frac{me^4}{8\epsilon_0^2 h^3}\left[\frac{1}{(n-1)^2} - \frac{1}{n^2}\right]$$

$$= \frac{me^4}{8\epsilon_0^2 h^3}\cdot\frac{(2n-1)}{n^2(n-1)^2}$$

For large n (i.e. $n \gg 1$); $(2n - 1) \simeq 2n$, $(n - 1)^2 \simeq n^2$ so that, we get

$$\nu = \frac{me^4}{4\epsilon_0^2 h^3}\cdot\frac{1}{n^3} \qquad \text{(i)}$$

The frequency at which the electron revolves round the nucleus is given by [see Eq. (13.12)].

$$\nu_r = \frac{v_n}{2\pi r_n}$$

Using Eq. (13.9) for r_n and Eq. (13.11) for v_n, we get

$$\nu_r = \frac{me^4}{4\epsilon_0^2 h^3}\cdot\frac{1}{n^3} \qquad \text{(ii)}$$

which is the same the expression for ν in Eq. (i) above.

EXAMPLE 13.4 Using Bohr's formula for energy quantization, determine

(i) the largest wavelength in the Balmer series of the hydrogen atom spectrum,

(ii) the excitation energy of the $n = 3$ level of He^+ atom.

(iii) the ionization potential of the ground state of Li^{++} atom.

Solution: (i) For Balmer series, $n_2 = 2$ so that, from Eq. (13.15) the frequencies of radiation for the Balmer series are given by [setting $n_2 = 2$ and $n_1 = n$ in Eq. (13.15)]

$$\nu_n = \frac{me^4}{8\epsilon_0^2 h^3}\left(\frac{1}{4} - \frac{1}{n^2}\right) \qquad \text{(i)}$$

Hence $$\text{Wavelength } \lambda_n = \frac{c}{\nu_n} \qquad \text{(ii)}$$

In Balmer series, the smallest frequency corresponds to $n = 3$ which is

$$\nu_3 = \frac{me^4}{8\epsilon_0^2 h^3}\left(\frac{1}{4} - \frac{1}{9}\right)$$

$$= 4.571 \times 10^{14} \text{ Hz}$$

Using this value in (ii) we get ($\because c = 3.0 \times 10^8 \text{ ms}^{-1}$)

$$\lambda_3 = 6563 \text{ Å}$$

(ii) For He^+ atom, the atomic number $Z = 2$. The energy required to excite the atom from the ground state ($n = 1$) to the $n = 3$ state is given by

$$E_3 - E_1 = \frac{m(Ze^2)^2}{8\epsilon_0^2 h^2}\left(\frac{1}{1^2} - \frac{1}{3^2}\right)$$

$$= 7.696 \times 10^{-18} \text{ J} = 48.1 \text{ eV}$$

(iii) For Li^{++} atom, $Z = 3$. The ionization energy is the energy required to shift the electron from the ground state ($n = 1$) to the state $n = \infty$. Thus

Ionization energy of Li^{++} is

$$E_\infty - E_1 = \frac{m(Ze^2)^2}{8\epsilon_0^2 h^2} \simeq 122 \text{ eV}$$

Therefore, the ionization potential is 122 V.

EXAMPLE 13.5 An electron of energy 20 eV collides with a hydrogen atom in the ground state. As a result of the collision, the atom is excited to a higher energy state and the electron is scattered with reduced velocity. The atom subsequently returns to its ground state with emission of radiation of wavelength 1.216×10^{-7} m. Find the velocity of the scattered electron.

Solution: The energy lost by the electron in exciting the hydrogen atom = energy corresponding to $\lambda = 1.216 \times 10^{-7}$ m,

$$= h\nu = \frac{hc}{\lambda}$$

$$= \frac{6.63 \times 10^{-34} \times 3.0 \times 10^{8}}{1.216 \times 10^{-7}} = 16.36 \times 10^{-19} \text{ J}$$

Now,

$$\text{Initial energy of electron} = 20 \text{ eV} = 32 \times 10^{-19} \text{ J}$$

Hence the kinetic energy of the scattered electron is

$$E = 32 \times 10^{-19} \text{ J} - 16.36 \times 10^{-19} \text{ J}$$

$$= 15.64 \times 10^{-19} \text{ J}$$

The velocity v of the scattered electron is given by

$$\frac{1}{2} m v^2 = E$$

or

$$v = \sqrt{\frac{2E}{m}}$$

$$= \left[\frac{2 \times 15.64 \times 10^{-19}}{9.11 \times 10^{-31}}\right]^{1/2}$$

$$= 1.86 \times 10^{6} \text{ ms}^{-1}$$

13.4 ATOMIC MASSES

As we shall see later, an accurate determination of the atomic mass of an element is essential because it provides considerable insight into the nulear phenomena. Mass spectrometers are instruments used to measure atomic masses. A variety of mass spectrometers have been devised by different workers but a simple spectrometer, called the Bainbridge mass spectrometer, will be discussed here. A schematic diagram of this spectrometer is shown in Fig. 13.9. It is based on the motion of charged particles in electric and magnetic fields. The substance or element whose atomic mass is to be determined is obtained in the form of ions. This can be done either by electron bombardment (if the substance is a gas) or by electric arc discharge (if the substance is a solid). A beam of ions, emerging from their source, is accelerated in an electric field (not shown in figure) and made to pass through two very narrow slits S_1 and S_2 which act as a collimator for the ion beam. A narrow beam of ions from S_1S_2 enters the velocity

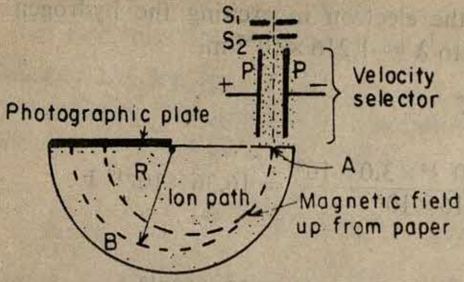


Fig. 13.9 *Bainbridge mass spectrometer*

selector. The velocity selector allows the ion beam of a particular velocity to pass through it and it consists of uniform electric and magnetic fields at right angles to each other and to the beam. The direction and magnitude of these fields are so adjusted that the deflection of the beam due to electric field counteracts the deflection due to magnetic field, i.e. the forces due to two fields on the ion are in opposite directions. If E and B are the magnitude of electric and magnetic fields, respectively, and q and v the charge and velocity of the ion, $qE = Bqv$ for no net deflection of ion in the velocity selector.

$$\therefore \qquad v = \frac{E}{B}$$

Only the ions having velocity equal to E/B pass through the velocity selector vertically while the others will hit the electric field plates PP. These ions of the same velocity enter a region of uniform magnetic field after emerging from a narrow opening A. In the magnetic field they follow circular paths whose radius R can be found by equating the magnetic force Bqv with centripetal force, i.e.

$$Bqv = \frac{mv^2}{R}$$

or

$$m = \frac{RBq}{v}$$

Since q, B and v are known, determination of R yields the value of m, the mass of the ion.

The ions after traversing a semicircular path in the magnetic field strike the photographic plate. Wherever the beam of ions strikes the plate, a dark line is produced on the plate. The distance between the opening A and the position of the dark line gives $2R$ (and hence R). Knowing R, m can be determined.

Atomic Mass Unit

The mass of the carbon-12 atom is measured to be 1.992678×10^{-26}kg which is very small. Therefore, a convenient unit is chosen (arbitrarily) for expressing the mass of atoms. This unit is defined to be 1/12 of the mass of carbon-12 atom, and its symbol is u. Thus

$$1 \text{ u} = \frac{\text{mass of carbon-12 atom}}{12}$$

$$= \frac{1.992678 \times 10^{-26} \text{ kg}}{12}$$

or

$$1 \text{ u} = 1.660565 \times 10^{-27} \text{ kg}$$

In terms of this unit, the mass of an electron (m_e) = 0.00055 u = 9.11×10^{-31} kg and the mass of a proton (m_p) = 1.0073 u = 1.6726×10^{-27} kg. The mass of a hydrogen atom = $m_e + m_p$ = 1.0078 u. Table 13.2 shows the masses of some common elements in terms of the atomic mass unit (u) defined above.

Table 13.2 Atomic masses of some common elements

Atomic number	Element	Symbol	Atomic mass (u)
1	Hydrogen	H	1.00794
2	Helium	He	4.0026
6	Carbon	C	12.011
7	Nitrogen	N	14.0067
8	Oxygen	O	15.9994
11	Sodium	Na	22.9898
17	Chlorine	Cl	35.453
19	Potassium	K	39.0983
20	Calcium	Ca	40.08
26	Iron	Fe	55.847
29	Copper	Cu	63.546
47	Silver	Ag	107.868
78	Platinum	Pt	195.08
79	Gold	Au	196.966
80	Mercury	Hg	200.59
82	Lead	Pb	207.2
92	Uranium	U	238.0289

13.5 DISCOVERY OF THE NEUTRON

According to the Rutherford and Bohr model, the mass of any atom is nearly equal to the mass of its nucleus which contains positively charged particles called protons. The number of protons is equal to the number of electrons and this is called the atomic number of the atom. If one looks at Table 13.2, one finds that the atomic mass is approximately twice the mass of protons in case of light nuclei and this ratio is more in the case of heavier nuclei. For example, helium has a atomic mass of 4.0026 u and it has an atomic number 2, i.e. it has two electrons and two protons. The mass of two electrons and two protons comes out to be $\simeq 3.34 \times 10^{-27}$ kg or 2.01565 u, which is nearly half the measured mass of helium. Similarly, one can take the case of any other element. Many other hypotheses were put forward to explain this anomaly, but were rejected for one reason or the other. This discrepancy of the mass of the nucleus continued till the discovery of the neutron by James Chadwick in 1932. It was found earlier that a highly penetrating radiation is produced when an alpha particle (helium nucleus) strikes a light metal such as beryllium. Chadwick performed additional experiments and established that the penetrating radiations consist of particles rather than electromagnetic waves. Further, he found that (i) the mass of these particles is nearly equal to that of proton and (ii) they are electrically neutral because of their high penetrating power. He gave the name *neutron* to this new particle. Neutrons account not only for the experimental results described above but also fit perfectly well to account for observed atomic masses of elements.

Chadwick's Estimate of the Mass of a Neutron

Chadwick estimated the mass of a neutron by comparing the maximum velocity U_H imparted to a hydrogen atom when a neutron collides with it head-on with the maximum velocity U_N imparted to a nitrogen atom when a neutron has a head-on collision with it. Let m be the mass of a neutron and V its initial velocity. Let M be the mass of the stationary atom (hydrogen or nitrogen) and let U be its speed after a head-on collision with the neutron. Let V' be the speed of the neutron after the head-on collision. Then from the conservation of linear momentum, we have

$$mV = MU + mV' \qquad (13.18)$$

The conservation of energy gives

$$\frac{1}{2}mV^2 = \frac{1}{2}MU^2 + \frac{1}{2}mV'^2 \qquad (13.19)$$

Eliminating V' from Eqs (13.18) and (13.19), we have

$$U = \frac{2mV}{(m+M)} \qquad (13.20)$$

The speeds U_H and U_N are obtained from Eq. (13.20) by substituting $M = 1$ unit for hydrogen and $M = 14$ units for nitrogen respectively, i.e.

$$U_H = \frac{2mV}{(m+1)}$$

and
$$U_N = \frac{2mV}{(m+14)}$$

giving
$$\frac{U_H}{U_N} = \frac{(m+14)}{(m+1)} \qquad (13.21)$$

Chadwick measured U_H and U_N. His values were $U_H = 3.7 \times 10^7 \text{ms}^{-1}$ and $U_N = 4.7 \times 10^7 \text{ ms}^{-1}$. Substituting these measured values in Eq. (13.21), Chadwick estimated $m \simeq 0.9$ u. The mass of a neutron is now known to a high degree of accuracy. The modern accepted value is

$$m_n = 1.00866 \text{ u} = 1.6749 \times 10^{-27} \text{ kg}$$

Chadwick was awarded the Nobel prize in physics in 1935 for his discovery of the neutron.

Composition of the Nucleus

The composition of a nucleus can now be described as follows: the nucleus of an atom contains *protons* and *neutrons*; they are collectively called *nucleons*. The total number of nucleons is called *mass number* and is denoted by A. The number of protons in a nucleus is called its *atomic number* which is denoted by Z. The number of neutrons is denoted by N, so that

$$A = Z + N$$

i.e. mass number = total number of protons + total number of neutrons = total number of nucleons.

According the latest convention, nuclear species are described by the notation

$${}^{A}_{Z}X$$

where X is the chemical symbol of the element whose mass number is A and atomic number Z. For example, helium nucleus has two protons and two neutrons and the chemical symbol for helium He, so a helium nucleus is written as ${}^{4}_{2}\text{He}$. Similarly ${}^{238}_{92}\text{U}$ represents a uranium nucleus consisting of 92 protons and (238–92) = 146 neutrons.

Isotopes

After a careful analysis of the photographic plate of a mass spectrometer (Fig. 13.9), it was found that the ions of the same element give more than one line, out of which one is much darker than the rest. This shows that the ions, and hence, atoms of the same element have more than one mass out of which one is relatively more abundant than the others. The atomic mass of a given element is the weighted average of these masses. For example, when a study of chlorine was made, two lines, which corresponded to masses 34.980 u and 36.977 u, were observed. Also, chlorine was found to be composed of 75% of the former and 25% of the latter. The weighted average of the mass of chlorine is 35.453 u. Also, the former corresponds to a mass number of 35 while the latter corresponds to 37. Such atoms of the same element are called isotopes. Isotopes are atoms of the same element which have different atomic masses, i.e. they have the same *Z* but different *A*. Since the atomic number of chlorine is 17, the two isotopes of chlorine are $^{35}_{17}Cl$ and $^{37}_{17}Cl$. Similarly, oxygen has three isotopes $^{16}_{8}O$, $^{17}_{8}O$ and $^{18}_{8}O$.

13.6 MASS DEFECT AND BINDING ENERGY

It can be seen easily that the mass of the nucleus of an atom is not equal to the sum of the masses of its nucleons. For example, let us consider the $^{4}_{2}He$ nucleus. Helium nucleus has 2 neutrons and 2 protons and the mass of helium nucleus is 4.0015 u.

$$\text{Mass of two protons} = 2\times 1.0073 = 2.0146 \text{ u}$$

$$\text{Mass of two neutrons} = 2\times 1.0087 = 2.0174 \text{ u}$$

$$\therefore \quad \text{Total mass} = 4.0320 \text{ u}$$

Therefore, the mass of the constituent parts of the helium nucleus is greater than the mass of the helium nucleus. This holds for all nuclei. This difference between the rest mass of a nucleus and the sum of the rest masses of its constituent nucleons is called *mass defect*. In the case of helium,

$$\text{Mass defect} = 4.0320 - 4.0015 = 0.0305 \text{ u}$$

Similarly, the mass of deuteron ($^{2}_{1}H$) nucleus, an isotope of hydrogen, is 2.0136 u, and the combined mass of its constituent nucleons (one proton and one neutron) is 2.0160 u.
Therefore,

$$\text{Mass defect} = 2.0160 - 2.0136 = 0.0024 \text{ u}$$

The reason for the mass defect in a nucleus can be understood in terms of Einstein's famous *mass energy equivalence* relation:

$$E = mc^2 \tag{13.22}$$

where E is the energy of an object of mass m and c is the speed of light in free space. It states that an object of mass m has an energy content equal to its mass times the square of the speed of light. Thus an object of mass 1 kg has an energy content of $(1\ \text{kg}) \times (3.0 \times 10^8\ \text{ms}^{-1})^2 = 9 \times 10^{16}$ J. Equivalently, the mass associated with an energy E is E/c^2.

In units of MeV (million electron volt), the energy equivalent of the mass of an elcctron $= m_e c^2 = (9.11 \times 10^{-31}\ \text{kg}) \times (3.0 \times 10^8\ \text{ms}^{-1})^2 = 8.2 \times 10^{-14}\ \text{J} = 0.511 \times 10^6\ \text{eV} = 0.511$ MeV. Thus

$$1\ \text{u} = 931.5\ \text{MeV}$$

Binding Energy

When the nucleons combine to form the nucleus the total energy of the system (including the mass energy equivalence) remains constant although the mass may not. In the nucleus, the nucleons are held or bound together with some potential energy called the *binding energy* which comes at the expense of the increase in the total mass of the system. Since the energy is conserved, the decrease in mass (according to Einstein's mass-energy equivalence) or the energy of the system, means that energy must be released in combining the nucleons together in the nucleus. Correspondingly, the same energy must be given to a nucleus if it is desired to break the nucleus into its parts. This situation is analogous to the situation in chemical reactions. We know that energy is released when carbon and oxygen combine to form carbon dioxide and the same amount of energy has to be added if we wish to decompose carbon dioxide into carbon and oxygen. In the case of helium, the energy released would be $0.03054\ \text{u} = 0.03054 \times 931 = 28.3$ MeV, while in deuteron it would be $0.0024\ \text{u} = 2.22$ MeV. If we wish to break a deuteron nucleus into a proton and a neutron, a photon of 2.22 MeV can be used as shown in Fig. 13.10.

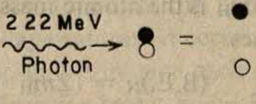


Fig. 13.10 *Breaking a deuteron by a photon*

The total energy that would be released if a nucleus is built from its constituents is called the binding energy of the nucleus. In the case of deuteron, it is 2.22 MeV, while in the case of $^4_2$He it is 28.3 MeV. These values are about a million times greater than the energy of the atomic electrons. The binding energy can also be defined in the reverse way, i.e. the energy required to separate a nucleus into its constituents.

Consider a nucleus ${}^{A}_{Z}\mathrm{X}$ of atom X; it contains Z protons and $(A-Z)$ neutrons. The mass defect is

$$\Delta m = Zm_{\mathrm{p}} + (A - Z)m_{\mathrm{n}} - m_{\mathrm{N}}$$

where m_{N} is the nuclear mass (not the atomic mass) of ${}^{A}_{Z}\mathrm{X}$. Therefore, the binding energy of the nucleus is (using mass-energy equivalence)

$$(\mathrm{B.E.})_{\mathrm{N}} = (\Delta m)c^2 = [Zm_{\mathrm{p}} + (A - Z)m_{\mathrm{n}} - m_{\mathrm{N}}]c^2 \qquad (13.23)$$

This is the energy needed to split the nucleus into Z protons and $(A - Z)$ neutrons. Now the mass of the nucleus (m_{N}) is not known. What is known (i.e. measured by experiment) is the atomic mass m which includes the mass of Z electrons also (i.e. Zm_{e}). These electrons are bound to the nucleus by electrostatic forces and, therefore, have a binding energy of their own, which from the mass-energy equivalence is given by

$$(\mathrm{B.E.})_{\mathrm{e}} = [(m_{\mathrm{N}} + Zm_{\mathrm{e}}) - m]c^2 \qquad (13.24)$$

Now the binding energy of electrons $(\mathrm{B.E.})_{\mathrm{e}}$ is negligible compared to the binding energy of the nucleons $(\mathrm{B.E.})_{\mathrm{N}}$; the former is of the order of eV to keV and the latter is of the order of 10^3 MeV. Hence it is a safe approximation to set $(\mathrm{B.E.})_{\mathrm{e}} = 0$ in Eq. (13.24) so that we get

$$m_{\mathrm{N}} + Zm_{\mathrm{e}} - m = 0$$

or

$$m_{\mathrm{N}} = m - Zm_{\mathrm{e}}$$

Thus, in terms of atomic mass m, Eq. (13.23) reads

$$(\mathrm{B.E.})_{\mathrm{N}} = [Z(m_{\mathrm{p}} + m_{\mathrm{e}}) + (A - Z)m_{\mathrm{n}} - m]c^2 \qquad (13.25)$$

Now, we know that hydrogen atom contains one proton and one electron, so that

$$m_{\mathrm{H}} = m_{\mathrm{p}} + m_{\mathrm{e}}$$

where m_{H} is the atomic mass of hydrogen. In terms of m_{H}, Eq. (13.25) becomes

$$(\mathrm{B.E.})_{\mathrm{N}} = [Zm_{\mathrm{H}} + (A - Z)m_{\mathrm{n}} - m]c^2 \qquad (13.26)$$

Therefore, binding energy per nucleon is

$$\bar{B} = \frac{(\mathrm{B.E.})_{\mathrm{N}}}{A} = \frac{[Zm_{\mathrm{H}} + (A - Z)m_{\mathrm{n}} - m]c^2}{A} \qquad (13.27)$$

The binding energy per nucleon can be computed for all nuclei. The plot of the binding energy per nucleon versus the mass number A is

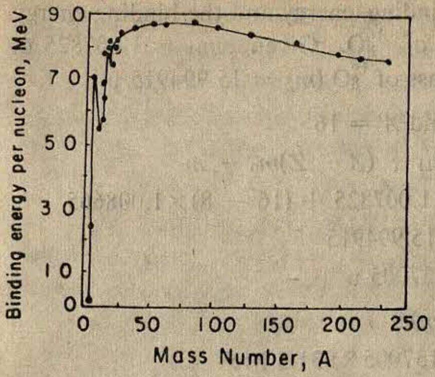


Fig. 13.11 *Plot of B as a function of A*

interesting (Fig. 13.11). Except for an anomalously high peak near $^4_2$He, the curve is regular. The nuclei in the middle range (A from 50 to 80) have the highest binding energy per nucleon showing thereby that the elements ranging from A = 50 to A = 80 are the most stable. Another significant feature of this curve is that as one approaches the central portion of the curve, i.e. the region of higher binding energy per nucleon, energy is liberated. As we shall discuss later, when thermal neutrons strike a $^{235}_{92}$U atom, uranium breaks up into two light elements and the binding energy per nucleon changes from 7.6 MeV to nearly 8.4 MeV. Greater binding energy per nucleon results in the liberation of energy. This is what happens in fission which is the basis of the atom bomb. Similarly, when two extremely light nuclei are combined to form a heavier one, the higher binding energy per nucleon of the latter also results in the evolution of energy; e.g. when two deuterons are joined to make a helium ($^4_2$He) nucleus, about 23 MeV would be liberated. This is what happens in fusion which is the basis of the hydrogen bomb.

The increase in the binding energy per nucleon for very light nuclei can be attributed to the fact that as the mass number increases, on the average, each nucleon has more neighbours to which it is bound. But this increase is not steady and at some stage saturation occurs, i.e. on the average, the same number of nucleons surround each nucleon. The decrease in binding energy per nucleon for heavy nuclei is due to large Coulomb repulsion between a proton in the outer part of the nucleus and all other protons. It is easy to show that in the case of a $^{235}_{92}$U nucleus (radius $\sim 9 \times 10^{-15}$ m), the Coulomb repulsive force between an outer proton and the rest of the 91 protons is about as large as 270 N.

EXAMPLE 13.6 Determine the binding energy and the binding energy per nucleon of the nucleus of ${}^{16}_{8}O$. Given $m_H = 1.007825$ u, $m_n = 1.008665$ u and atomic mass of ${}^{16}_{8}O$ $(m) = 15.994915$ u.

Solution: In this case, $Z = 8$ and $A = 16$

$$\therefore \quad \text{Mass defect } (\Delta m) = Zm_H + (A - Z)m_n - m$$

$$= 8 \times 1.007825 + (16 - 8) \times 1.008665 - 15.994915$$

$$= 0.137005 \text{ u}$$

$$\therefore \quad \text{Binding energy} = (\Delta m)c^2$$

$$= 0.137005 \times 931.5 \text{ MeV}$$

$$= 127.620 \text{ MeV}$$

$$\text{Binding energy per nucleon} = \frac{127.620 \text{ MeV}}{16}$$

$$= 7.976 \text{ MeV}$$

EXAMPLE 13.7 How much energy is needed to remove a neutron from the nucleus of ${}^{41}_{20}Ca$? Given $m_n = 1.008665$ u, atomic mass of ${}^{41}_{20}Ca$ $(m) = 40.962278$ u and atomic mass of ${}^{40}_{20}Ca$ $(m) = 39.962591$ u.

Solution: Let m'_N be the nuclear mass of ${}^{40}_{20}Ca$ and m_N that of ${}^{41}_{20}Ca$. Then the neutron separation energy S_n (i.e. the energy needed to remove a neutron from the nucleus) is given by

$$S_n = (m'_N + m_n - m_N)c^2$$

$$= [(m'_N + Zm_e) + m_n - (m_N + Zm_e)]c^2$$

(adding and subtracting Zm_e).

But $m'_N + Zm_e = m'$ and $m_N + Zm_e = m$, where we have ignored the electron binding energies as before. Hence, in terms at atomic masses m and m', we have

$$S_n = (m' + m_n - m)\, c^2$$

Using the given values of m, m' and m_n and noting that $c^2 = 931.5$ MeV/u, we get

$$S_n = 8.362 \text{ MeV}$$

13.7 NUCLEAR DIMENSIONS AND NUCLEAR DENSITY

Rutherford's experiments of the scattering of alpha particles from thin metal foils provide a rough estimate of the size of the nucleus ($\sim 10^{-15}$ m). By performing scattering experiments with fast electrons, sizes of nuclei of different elements have been accurately measured. If the nuclei are assume spherical, their volumes can be estimated. If R is the radius of the sphere, its volume is $4\pi R^3/3$.

Experimental measurement show that the volume of a nucleus is proportional to its mass number A. This means that $R \propto A^{1/3}$ or

$$R = R_0 A^{1/3} \tag{13.28}$$

where the empirical constant R_0 is found to be 1.1×10^{-15}m. Since volume $\propto R^3$ and $R^3 \propto A$, it is clear that the density of nuclear matter is independent of the size of the nucleus. For a helium nucleus, $A = 4$. Therefore the radius of helium nucleus is

$$R = R_0 A^{1/3} = 1.1 \times 10^{-15} \times (4)^{1/3} \simeq 1.9 \times 10^{-15} \text{ m}$$

$$\therefore \quad \text{Volume of helium nucleus} = \frac{4\pi R^3}{3} \simeq 28.9 \times 10^{-45} \text{ m}^3$$

Also the mass of a helium nucleus = 4.0015 u = 6.64×10^{-27} kg.

Hence, the density of helium nucleus is

$$\rho = \frac{6.64 \times 10^{-27}}{28.9 \times 10^{-45}} \simeq 2.29 \times 10^{17} \text{ kg m}^{-3}$$

It is very large when compared to the density of ordinary matter, say water, for which $\rho = 10^3$ kg m^{-3}. Thus nuclear matter has enormous density of the order of 10^{17} kg m^{-3}. Such high densities are found in *white dwarf stars* or *neutron stars* in the universe (see Chapter 15).

EXAMPLE 13.8 Find the approximate ratio of the nuclear radii of iron $^{56}_{26}$Fe and calcium $^{40}_{20}$Ca. What is the approximate ratio of their nuclear mass densities?

Solution:

$$R(\text{Fe}) = R_0 \times (56)^{1/3} \qquad \text{and} \qquad R(\text{Ca}) = R_0 \times (40)^{1/3}$$

$$\therefore \quad \frac{R(\text{Fe})}{R(\text{Ca})} = \left(\frac{56}{40}\right)^{1/3} = 1.12$$

The density of nuclear matter is almost independent of the radius of the nucleus. Thus, the nuclear mass densities of Fe and Ca are very nearly equal.

13.8 NUCLEAR FORCES

An interesting question arises when we consider the fact that the protons in a nucleus should repel each other since they are each positively charged. Experiments tell us that the average distance between protons is approximately 10^{-15} m. Coulomb's law tells us that at such short distances the repulsive forces between protons must be extremely large.

Why don't the protons fly off from the nucleus under such a large repulsive force? What holds them together in the nucleus? These questions baffled the scientists for quite some time. They argued that the gravitational force of attraction between protons is much weaker than the electrostatic repulsive force and hence could not hold the nucleus together. Some new force that is much stronger than the previously known forces must be involved. They called this new force the *nuclear force.*

The nuclear force is an attractive force that acts between all nucleons. It is a very strong force; but we do not feel it in everyday life. The reason is that nuclear force is a very *short-range* force. It is very strong between two nucleons which are 10^{-15} m apart and is almost negligible at a distance only slightly greater than this. On the other hand, the electrostatic, magnetic and gravitational forces are long-range forces and so we can feel them even at fairly large distances.

Nuclear forces have a strange behaviour. If a nucleus contains too many protons compared to the number of neutrons, or too many neutrons compared to the number of protons, the nuclear force becomes weak. The nuclei of such elements become unstable and show a tendency to disintegrate. Nuclei with $Z > 82$ (i.e. lead) show this unstable behaviour. They acquire stability by emitting alpha, beta and gamma rays. This phenomenon of self-emission is called radioactivity which we will study in the next section.

Another feature of nuclear forces is that they are *charge-independent.* Experiments on the scattering of protons and of neutrons from nuclei show that the nuclear force between two protons or between two neutrons or between a proton and a neutron, is nearly equal showing that nuclear force does not depend on the charge of the particles.

13.9 RADIOACTIVITY

The story of the nucleus actually began in 1896 when Henri Becquerel accidently discovered natural radioactivity. One day he left some uranium salt on a photographic plate. The plate was wrapped in black paper and was lying in a dark room. When he developed the photographic plate, he found, to his surprise, that the plate had got affected and was darkened. Later, he repeated the experiment with other salts of uranium. From his experiments he concluded that uranium and its salts emit invisible radiations which can pass through paper, wood, glass etc., and affect the photographic plate. The phenomenon of self-emission of radiation is called *radioactivity* and the substances which emit these radiations are called radioactive substances.

A few years after Becquerel's discovery, Marie Curie and her husband Pierre Curie discovered the highly radioactive element radium. Further experiments revealed that many other naturally occurring compounds of heavy elements like thorium also emit these radiations. Today we know that all the naturally occurring elements with atomic number Z greater than 82 (i.e. heavier than lead) are radioactive. Some examples are radium, polonium, thorium, actinium, uranium, etc.

It was found that, in each case, the radioactivity was not affected by the strongest physical and chemical treatments. Excessive heating or cooling or the action of powerful chemical reagents had no effect on the activity of these radioactive elements. It became clear that the electrons orbiting the atom were not responsible for radioactivity because if they were, the activity should have been affected by chemical reactions. The source of radioactivity, therefore, must be deep inside the atom, far from the orbiting electrons. Indeed, a few years later, Ernest Rutherford found that the radioactivity was accompanied by the transformation of one element into another. It then became clear that radioactivity is the result of the disintegration of an unstable nucleus. The isotopes of certain elements are not stable under the action of the nuclear force. These unstable nuclei decay by emitting radiation or rays.

Alpha, Beta and Gamma Radiations

In 1898, Rutherford and others began to study the nature of the emitted radiations. The radiations from a radioactive material such as radium were made to pass through a strong magnetic field as shown in Fig. 13.12. The direction of the magnetic field is perpendicular to the plane of the paper and away from the reader. This is shown by crosses (×) in the diagram which represent the ends of the arrows pointing into the paper. Under the action of the magnetic field, the radiation separated into three distinct parts as shown in the figure. One part of the radiation was deflected to one side, another

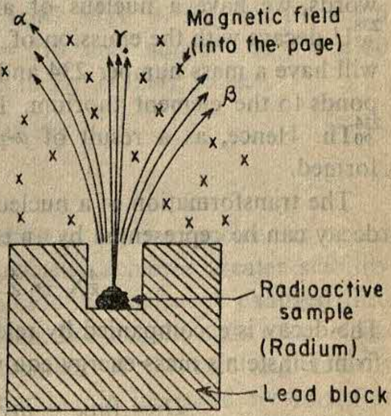


Fig. 13.12 *Effect of magnetic field on radiations emitted from a radioactive source*

part was deflected to the other side, while the third part remained undeflected. These three kinds of radiations were named alpha (α), beta (β) and gamma (γ) radiations.

Applying Fleming's left-hand rule, we find that the radiation that is deflected to the left (see Fig. 13.12) must have a positive charge; this radiation is called α-radiation or α-particles. The radiation that is deflected to the right has a negative charge and is called β-radiation or β-particles. The radiation that remains undeflected must be neutral and is called γ-radiation or γ-rays

An alpha particle is a helium nucleus ($^4_2$He) having two protons and two neutrons. It has a positive charge equal to the charge of 2 protons and mass about 4.0015 u. It has an initial speed of about 10^7 ms^{-1}. These particles have very little penetrating power and can be stopped even with a thin paper.

Beta particles are just electrons. Their speeds are nearly equal to the speed of light. They are much more penetrating than α-particles.

Gamma-rays are high frequency (or short-wavelength) electromagnetic radiations having a high-penetrating power.

Alpha Decay

The process of emission of an α-particle from a nucleus is called *alpha decay*. It occurs generally for nuclei containing 210 or more nucleons. When a nucleus emits an alpha particle, it loses two protons and two neutrons which means that the daughter nucleus has its mass number smaller by 4 and its atomic number smaller by 2. In other words, we have a nucleus of another element. For example, when $^{238}_{92}$U decays with the emission of an α-particle, the daughter nucleus will have a mass number 234 and an atomic number 90 which corresponds to the element thorium. Hence the resulting nucleus will be $^{234}_{90}$Th. Hence, as a result of α-particle emission, a new element is formed.

The transformation of a nucleus ^A_ZX into a nucleus $^{A-4}_{Z-2}$Y by alpha decay can be represented by an equation

$$^A_Z\text{X} \rightarrow {}^{A-4}_{Z-2}\text{Y} + {}^4_2\text{He} \tag{13.29}$$

The decay is accompanied by release of energy Q which can be found from Einstein's mass-energy equivalence relation which gives

$$Q = (m_\text{X} - m_\text{Y} - m_\text{He})\, c^2 \tag{13.30}$$

The daughter nucleus $^{A-4}_{Z-2}$Y and the alpha particle $^4_2$He share this energy.

The speed of the emitted alpha-particle can be calculated by applying the principles of conservation of momentum and energy. Suppose the parent nucleus ${}^{A}_{Z}\mathrm{X}$ is at rest before it undergoes alpha decay. Let v_{He} and v_{Y} be the velocities of the alpha particle and the daughter nucleus. From the principle of conservation of mometum, we have

$$m_{\mathrm{Y}}v_{\mathrm{Y}} = m_{\mathrm{He}}v_{\mathrm{He}} \tag{13.31}$$

Let Q be the energy released in alpha decay. From the principle of conservation of energy, this energy must be equal to the sum of the kinetic energies of the nucleus Y and the alpha particle, i.e.

$$Q = \frac{1}{2}\, m_{\mathrm{Y}}\, v_{\mathrm{Y}}^{2} + \frac{1}{2}\, m_{\mathrm{He}}\, v_{\mathrm{He}}^{2} \tag{13.32}$$

Eliminating v_{Y} from Eqs (13.31) and (13.32), we get

$$\frac{1}{2}\, m_{\mathrm{He}}\, v_{\mathrm{He}}^{2} = \frac{m_{\mathrm{Y}}}{(m_{\mathrm{Y}} + m_{\mathrm{He}})}\cdot Q \tag{13.33}$$

Now $m_{\mathrm{Y}} \simeq (A - 4)$ atomic mass units and $m_{\mathrm{He}} \simeq 4$ atomic mass units. Using these in Eq. (13.33), we have

$$\frac{1}{2}\, m_{\mathrm{He}}\, v_{\mathrm{He}}^{2} \simeq \frac{A-4}{A}\cdot Q$$

or

$$v_{\mathrm{He}} \simeq \sqrt{\frac{2(A-4)Q}{A\, m_{\mathrm{He}}}} \tag{13.34}$$

where the value of Q is computed from Eq. (13.30).

Let us compute v_{He} in the decay of a radon nucleus ${}^{222}_{86}\mathrm{Rn}$. In this, α-decay the energy released is $Q = 5.587$ MeV. $= 5.587 \times 10^{6} \times 1.6$ 10^{-19} J. Also $A = 222$, $(A - 4) = 218$ and $m_{\mathrm{He}} \simeq 4.00\ \mathrm{u} = 4.00 \times 1.66 \times 10^{-27}$ kg. Using these values in Eq. (13.34), we get

$$v_{\mathrm{He}} \simeq 1.62 \times 10^{7}\ \mathrm{ms}^{-1}$$

Beta Decay

The process of the emission of an electron from a nucleus is called *beta decay*. In this process also, the nucleus achieves greater stability by emitting an electron. A neutron inside the nucleus is changed into a proton by the emission of an electron. Because the mass of an electron is negligibly small, the mass number of the resulting nucleus remains unaltered but its atomic number is increased by one. For example, when a radium nucleus ${}^{228}_{88}\mathrm{Ra}$ emits a β-particle, the resulting element is an isotope of actinium ${}^{228}_{89}\mathrm{Ac}$. Thus in β-decay also, a new element is formed.

The transformation of a nucleus ${}_{Z}^{A}\mathrm{X}$ into the nucleus ${}_{Z+1}^{A}\mathrm{Y}$ by β-decay is represented by an equation

$$ {}_{Z}^{A}\mathrm{X} \rightarrow {}_{Z+1}^{A}\mathrm{Y} + e^- \tag{13.35} $$

For a given disintegrating nucleus, the energy (and hence velocity) of the α-particle in alpha decay is fixed [see Eq. (13.33)]. But this is not true of β-decay. Experiments show that the energy of the emitted electron varies continuously from zero to a maximum value. To account for this, Pauli, in 1930, postulated that an uncharged particle (which interacts very weakly with matter and hence escapes undetected) is emitted along with the electron. This particle was named *antinutrino* ($\bar{\nu}$) so that Eq. (13.35) is to be corrected as follows:

$$ {}_{Z}^{A}\mathrm{X} \rightarrow {}_{Z+1}^{A}\mathrm{Y} + e^- + \bar{\nu} \tag{13.36} $$

Similarly the β-decay of a free neutron (n) into a proton (p) is correctly described as

$$ \mathrm{n} \rightarrow \mathrm{p} + \mathrm{e}^- + \bar{\nu} \tag{13.37} $$

Gamma Decay

Gamma rays are high-frequency electromagnetic radiations (i.e. photons) which do not carry any charge. Hence in γ-decay, the mass number and the atomic number of the nucleus remain unchanged so that no new element is formed.

When an α-decay or a β-decay occurs in a radio-isotope, the daughter nucleus is mostly in an excited state and it achieves stability by the emission of one or more gamma rays (photons). For example, the unstable cobalt nucleus ${}_{27}^{60}\mathrm{Co}$ β-decays to an excited ${}_{28}^{60}\mathrm{Ni}$ nucleus which in turn reaches the stable ground state by emitting photons of energies 1.17 MeV and 1.33 MeV in two γ-decay transitions.

Radioactive Decay Law

Since the number of atoms in any substance is very large, radioactive emissions are essentially statistical in nature, i.e. it is not possible to predict which particular nucleus will undergo radioactive decay in a given time interval. Yet it is possible to predict the decay behaviour of a collection of a large number of nuclei very accurately using statistical arguments.

Let λ be the *probability per unit time* for a nucleus of a radioactive sample to decay. The parameter λ depends upon the nature of the radioactive substance. Let N be the number of the radioactive nuclei present in the sample at an instant of time. The number dN of radioactive decays in a time interval dt can be obtained as follows. Since λ is the probability per unit time for a single nucleus to decay, in a time interval dt, the probability for that nucleus to decay will be λdt. Since number of radioactive nuclei present is N, the number $|\,\mathrm{d}N\,|$ of radioactive decays in time interval dt will be given by

$$\mathrm{d}N = -\lambda N\,\mathrm{d}t$$

The negative sign is included because the number N of radioactive nuclei decreases with time due to decays. Dividing by N we get

$$\frac{\mathrm{d}N}{N} = -\lambda \mathrm{d}t \tag{13.38}$$

Integrating this equation, we have

$$\ln N = -\lambda t + \ln A \tag{13.39}$$

where A is the integration constant. Let N_0 be the number of radioactive nuclei present initially, i.e. at time $t = 0$. Putting $t = 0$ and $N = N_0$ in Eq. (13.39), we get

$$\ln N_0 = \ln A$$

so that Eq. (13.39) becomes

$$\ln N = -\lambda t + \ln N_0$$

or

$$\ln\left(\frac{N}{N_0}\right) = -\lambda t$$

or

$$\frac{N}{N_0} = e^{-\lambda t}$$

or

$$N = N_0\, e^{-\lambda t} \tag{13.40}$$

Thus, the number of active nuclei in a radioactive sample decreases exponentially with time. Figure 13.13 shows a plot of N versus t which is the radioactive decay curve.

Half-life ($T_{1/2}$) Radioactive decays are characterized by a time constant $T_{1/2}$, called the *half-life*. The half-life of a radioactive isotope is the time in which half of the number of nuclei decay. For a given isotope, the value of half-life is a constant independent of N. This can be seen from the decay curve in Fig. 13.13. For example, the half-life of radium -226 ($^{226}_{88}$Ra) is 1620 years. This means that 2g

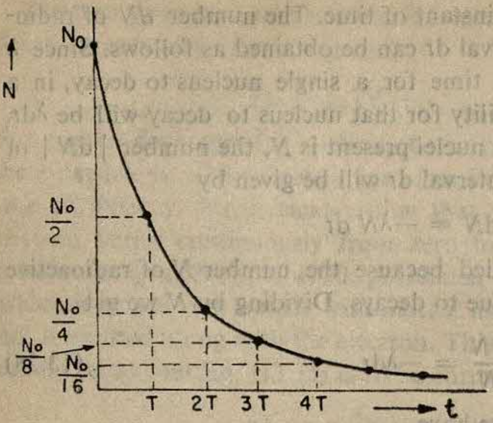


Fig. 13.13 *Radioactive decay curve. After $T_{1/2}$ seconds the number of unstable nuclei drops by a factor of 2*

of this isotope will decay to 1 g in 1620 years, to 0.5 g in 2 × 1620 years, and so on.

The value of $T_{1/2}$ is related to λ as follows. By definition, N given by Eq. (13.40) becomes equal to $N_0/2$ in a time $t = T_{1/2}$. Setting $N = N_0/2$ and $t = T_{1/2}$ in Eq. (13.40), we have

$$\frac{N_0}{2} = N_0\, e^{-\lambda T_{1/2}}$$

or

$$\frac{1}{2} = e^{-\lambda T_{1/2}}$$

Taking natural logarithm, we get

$$T_{1/2} = \frac{\ln 2}{\lambda} = \frac{0.693}{\lambda} \tag{13.41}$$

Using the definition of half-life, we can calculate the number of active nuclei left after a time t from the relation

$$N = N_0\left(\frac{1}{2}\right)^n \tag{13.42}$$

where n is the number of half-lives in time t, i.e. $n = t/T_{1/2}$.

The half-lives of radioactive nuclei vary over a very wide range from 10^{-15} s to 10^{10} years. The half-life of $^{212}_{84}Po$ is 3×10^{-7} s while that of $^{204}_{82}Pb$ is 1.4×10^7 years.

Radioactive Decay Rate It is useful to use the concept of the decay rate R which is defined as the number of radioactive disintegrations taking place in a sample per second, which from Eq. (13.38) is given by

$$R = \frac{\mathrm{d}N}{\mathrm{d}t} = -\lambda N$$

or

$$|R| = \lambda N = \frac{0.693}{T_{1/2}} \cdot N \tag{13.43}$$

As N decreases exponentially with time, R will also decrease exponentially with time. The SI unit of the decay rate R is called *curie* (symbol Ci) in honour of Madame M S Curie (1867-1934). It is defined as the decay rate of 3.7×10^{10} disintegrations per second, i.e.

$$\begin{aligned} 1 \text{ Ci(curie)} &= 3.7 \times 10^{10} \text{ disintegrations/s} \\ 1\text{m Ci(millicurie)} &= 3.7 \times 10^{7} \text{ disintegrations/s} \\ 1\mu \text{ Ci(microcurie)} &= 3.7 \times 10^{4} \text{ disintegrations/s} \end{aligned}$$

Radio Carbon Dating

One of the most important and recent methods used by anthropologists and archaeologists to determine the age of ancient excavations is called *radiocarbon dating*. In the atmosphere there are a number of charged particles (of low and high energy) and stable isotopes. When the particles strike the stable isotopes, a number of radioisotopes are produced. One of these radioisotopes is ${}^{14}_{6}\mathrm{C}$ which is produced by the bombardment of atmospheric nitrogen with high energy neutrons. ${}^{14}_{6}\mathrm{C}$ is unstable and decays to nitrogen-14 with a half-life of about 5,600 years. This carbon-14 is incorporated into the atmospheric carbon dioxide molecules which are taken by plants because they breathe in carbon dioxide. Therefore, some of the radio carbon is taken by plants and hence by animals who eat these plants. Ultimately, the concentration of ${}^{14}_{6}\mathrm{C}$ in all living organisms is in equilibrium with the ${}^{14}_{6}\mathrm{C}$ in the atmosphere because ${}^{14}_{6}\mathrm{C}$ decays and ${}^{14}_{6}\mathrm{C}$ is also taken in through carbon dioxide by the plants and hence living organisms. When an organism dies, it stops taking carbon-14 from the atmosphere, i.e. the exchange of ${}^{14}_{6}\mathrm{C}$ with the atmosphere stops, and the activity or concentration of carbon-14 present in the organism decreases with time. By measuring the ratio of the concentration of ${}^{14}_{6}\mathrm{C}$ to ${}^{12}_{6}\mathrm{C}$ in any ancient organism, say a tree, and comparing it with the ratio in the corresponding living organism, say a tree which is not cut down or has not died, one can determine the date when the organism died. W.F. Libby, an American chemist, made an extensive study of this kind and was able to find out the age of a number of archaeological artifacts. He was awarded the Nobel Prize in chemistry in 1963 for his work in this field.

EXAMPLE 13.9 The nucleus of ${}^{230}_{90}\text{Th}$ is unstable against α-decay with a half-life of 7.6×10^3 years. Write down the equation of the decay and estimate the kinetic energy of the emitted α-particle from the following data:

$$m({}^{230}_{90}\text{Th}) = 230.0381 \text{ u}, \; m({}^{226}_{88}\text{Ra}) = 226.0254 \text{ u}$$

and $m({}^{4}_{2}\text{He}) = 4.0026$ u.

Solution: The equation of the decay is

$${}^{230}_{90}\text{Th} \rightarrow {}^{226}_{88}\text{Ra} + {}^{4}_{2}\text{He} + Q$$

The energy Q is given by [see Eq. (13.30)]

$$Q = [m(\text{Th}) - m(\text{Ra}) - m(\text{He})]c^2$$

Using the given data and $c^2 = 931.5$ MeV/u, we get

$$Q = 9.41 \text{ MeV}$$

This energy is shared by Ra and He. If the original nucleus Th is at rest, i.e. if the momentum of the system before α-decay is zero, the total momentum after the decay will also be zero. Thus Ra and He will have equal and opposite linear momenta.

or

$$m_{\text{He}}v_{\text{He}} = -m_{\text{Ra}}v_{\text{Ra}}$$

$$m^2_{\text{He}}v^2_{\text{He}} = m^2_{\text{Ra}}v^2_{\text{Ra}} \tag{i}$$

Now

$$\frac{\text{K.E.(He)}}{\text{K.E.(Ra)}} = \frac{\frac{1}{2} m_{\text{He}}v^2_{\text{He}}}{\frac{1}{2} m_{\text{Ra}}v^2_{\text{Ra}}}$$

$$= \left(\frac{m_{\text{He}}\, v_{\text{He}}}{m_{\text{Ra}}\, v_{\text{Ra}}}\right)^2 \times \left(\frac{m_{\text{Ra}}}{m_{\text{He}}}\right)$$

$$= \frac{m_{\text{Ra}}}{m_{\text{He}}} \quad \text{[use Eq. (i)]}$$

$$\simeq \frac{226}{4} \simeq 56.5 \tag{ii}$$

i.e. the kinetic energy of He is 56.5 times that of Ra, the total energy being

$$\text{K.E.(He)} + \text{K.E.(Ra)} = 9.41 \text{ MeV} \qquad \text{(iii)}$$

From (ii) and (iii) we have

$$\text{KE(Ra)} = 0.16 \text{ MeV} \quad \text{and} \quad \text{K.E.(He)} \simeq 9.25 \text{ MeV}$$

EXAMPLE 13.10 The half-life of $^{230}_{90}\text{Th}$ against α-decay is 7.6×10^3 years. How many disintegrations per second occur in 1 g of $^{230}_{90}\text{Th}$?

Solution:

$$\begin{aligned} T_{1/2} &= 7.6 \times 10^3 \text{ years} \\ &= 7.6 \times 10^3 \times 365 \times 24 \times 60 \times 60 \text{ s} \\ &= 2.40 \times 10^{11} \text{ s} \end{aligned}$$

Now 1 mole of a substance has a mass equal to its atomic mass expressed in grams. Hence the number of moles in 1g of Th-230 is

$$\frac{1 \text{ g}}{230 \text{ g mol}^{-1}} = 4.348 \times 10^{-3} \text{ mol}$$

Also,
Number of atoms in a mole of a substance = Avagadro number

$$= 6.025 \times 10^{23} \text{ atoms/mole}$$

Hence the number (N) of atoms in 1 g of Th-230 is

$$\begin{aligned} N &= 4.348 \times 10^{-3} \times 6.025 \times 10^{23} \\ &= 32.2 \times 10^{20} \text{ atoms} \end{aligned}$$

The decay rate R is given by [see Eq. (13.43)]

$$\begin{aligned} R &= \frac{0.693}{T_{1/2}} \cdot N \\ &= \frac{0.693 \times 32.2 \times 10^{20}}{2.40 \times 10^{11}} \\ &= 9.3 \times 10^9 \text{ disintegrations per second} \end{aligned}$$

EXAMPLE 13.11 1 gram of caesium-137 ($^{137}_{55}\text{Cs}$) decays by β-emission with a half life of 30 years. If the initial activity of caesium is 1.0 mCi and atomic mass of caesium is ~137 kg per k mol, what is

(a) the resulting isotope?
(b) the number of atoms left after 5 years?
(c) the activity of caesium after 5 years?

Solution: (a) Since the mass remains constant and the atomic number increases by +1 in β-emission, the resulting isotope has mass number 137 and atomic number 56. This corresponds to the isotope $^{137}_{56}$Ba.

(b) Initial mass of caesium = 1g

Now Atomic mass = 137 kg per k mol

The number of atoms in 1 k mol = 6.023×10^{26}

The number of caesium atoms present in 1g is

$$N_0 = \frac{6.023 \times 10^{26} \times 1}{137 \times 10^3} = 4.396 \times 10^{21}$$

$$T_{1/2} = 30 \text{ years}$$

$$t = 5 \text{ years}$$

$$\therefore \quad n = \frac{t}{T_{1/2}} = \frac{1}{6}$$

$\therefore$ N = number of atoms present after 5 years $= N_0 \left(\frac{1}{2}\right)^{1/6}$

$$= 3.907 \times 10^{21}$$

(c) Initial activity (R_0) = 1.0 mCi = $1.0 \times 3.7 \times 10^7$ disintegrations per second

$\therefore$ Activity after 5 years $(R) = 0.37 \times 10^8 \times \left(\frac{1}{2}\right)^{1/6} = 0.329 \times 10^8 \text{ s}^{-1}$

$$= 0.89 \text{ mCi}$$

EXAMPLE 13.12 The normal activity of a living matter containing carbon is found to be 15 decays per minute per gram of carbon. An archaeological specimen gives 6 decays per minute per gram of carbon. If the half-life of carbon is 5730 years, estimate the approximate age of the specimen.

Solution: Let t be the age of the specimen. The rate of decay is given by [see Eq. (13.43)]

$$R = \lambda N = \lambda N_0 e^{-\lambda t}$$

or

$$R = R_0 e^{-\lambda t}$$

or

$$\frac{R}{R_0} = e^{-\lambda t}$$

It is given that $R/R_0 = 6/15 = 2/5$, therefore

$$\frac{2}{5} = e^{-\lambda t}$$

Taking the logarithm of both sides, we have

$$\ln\left(\frac{2}{5}\right) = -\lambda t$$

or

$$t = -\frac{\ln(2/5)}{\lambda} = +\frac{\ln(5/2)}{\lambda}$$

But $\lambda = \ln 2/T_{1/2}$. Therefore,

$$t = \frac{\ln(5/2)}{\ln 2}\cdot T_{1/2}$$

$$= \frac{0.916}{0.693}\times 5730 \text{ years}$$

$$= 7576 \text{ years}$$

13.10 NUCLEAR FISSION

Knowing the fact that the emission of a β-particle increases the atomic number by one, Fermi and his co-workers (1934) attempted to produce the elements beyond uranium ($Z=92$) which at that time was the last element in the periodic table. They bombarded uranium with neutrons and found that β-particles with different half lives were emitted. Therefore, they concluded that the elements with $Z > 92$, i.e. the elements heavier than uranium, had been formed. Hahn and Strassmann made similar experiments in 1939. After the chemical analysis of the products they concluded that one of the product nuclei is barium and not a heavier element as predicted earlier. They concluded that the neutron bombardment can cause a uranium nucleus to break apart, producing two or more fragments of moderate and comparable sizes. This process was called *nuclear fission* (from the latin word *fissio* which means *cleaving*). Further they found that the reaction is much more pronounced with thermal neutrons. Only uranium-235 ($^{235}_{92}U$) undergoes this process of fission though naturally occurring uranium has 99.3% of $^{238}_{92}U$ and 0.7% of $^{235}_{92}U$. We shall see that in this process there is a decrease in the mass of the system, and hence a release of energy. Since this process can be started at will, it can be controlled and the energy liberated provides a good source of energy.

It was observed that when one thermal neutron strikes a uranium nucleus, three neutrons are emitted. In the reaction observed by Hahn *et al*, the product nuclei were $^{141}_{56}Ba$ and $^{92}_{36}Kr$. Therefore, the reaction can be written as

$$^{235}_{92}U + {}^{1}_{0}n \rightarrow {}^{141}_{56}Ba + {}^{92}_{36}Kr + 3\,{}^{1}_{0}n + Q$$

where Q is the energy of reaction which can be calculated from the values of rest masses of different nuclei. The calculation of Q is given below.

Initial masses		*Final masses*	
$^{235}_{92}U$	235.0439 u	$^{141}_{56}Ba$	140.9139 u
$^{1}_{0}n$	1.0087 u	$^{92}_{46}Kr$	91.8973 u
		$3^{1}_{0}n$	3.0261 u
	236.0526 u		235.8373 u

$$\therefore \quad \text{Decrease in mass} = 236.0526 - 235.8373 = 0.215 \text{ u}$$

$$\therefore \quad Q = 0.215 \times 931.5 \text{ MeV} = 200 \text{ MeV}$$

Therefore, when one atom of $^{235}_{92}U$ undergoes fission, 200 MeV of energy is released. If 1 g of naturally occurring uranium, which has about 10^{+19} atoms of $^{235}_{92}U$ undergoes fission the total energy released would be 200×10^{19} MeV $= 10^{8}$ J. This is a very large amount of energy. Though these calculations assume ideal circumstances, it is found that 1.0 kg of uranium delivers as much energy as the combustion of about 3000 tonnes of coal.

Chain Reaction

As mentioned before, when one uranium atom undergoes fission, it releases 3 neutrons (the average number is 2.42). If more than one of these neutrons is able to cause fission in the other $^{235}_{92}U$ nuclei, the number of neutrons would increase rapidly. Thus, a chain reaction can be set up (Fig. 13.14). The fission would proceed at an everincreasing rate and in a very short time the whole of $^{235}_{92}U$ would be transformed with the release of a large amount of energy. If such a chain reaction is not controlled, the large energy released can cause a violent explosion and destroy everything that comes in its way. This is the principle of the atom bomb. Further, if the amount of uranium is too small, the chain reaction can stop before it releases the amount of energy required for explosion. Therefore, if the chain reaction is to start, it is necessary that the mass of uranium be greater than some minimum mass called the *critical mass* or *critical size*.

Nuclear Reactor

The large amount of energy released in nuclear fission can be used for many useful purposes if the reaction is carried out under controlled

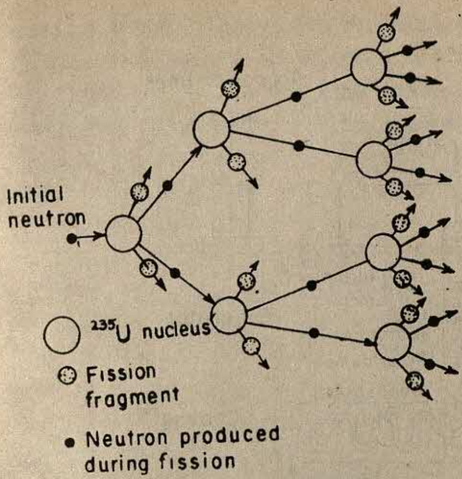


Fig. 13.14 *Illustrating a chain reaction in $^{235}_{92}$U*

conditions. A nuclear reactor is a device in which the chain reaction is a controlled one and the energy released can be used for any of the several purposes: to produce power, to supply neutrons, to prepare radioisotopes, etc. The first reactor was installed and operated by Fermi and his co-workers in 1942 in the USA. In reactors, small pieces of uranium are spread throughout a material, called the *moderator*, capable of slowing down the neutrons to thermal energies, so that they can cause fission in other nuclei. When a thermal neutron strikes a uranium atom, it starts the fission process which results in the splitting of the uranium atom and the production of more fast neutrons. These fast neutrons strike the material and lose their kinetic energy in repeated collisions with the nuclei of the material and get thermalized. These thermal neutrons strike another piece of uranium and again cause fission. Thus, a chain reaction is set up. Whenever this chain reaction is to be stopped, some material which is a strong absorber of neutrons is inserted in between so that the neutrons are absorbed and the rate of reaction is slowed down. Figure 13.15 shows the schematic diagram of a nuclear reactor. Basically, it consists of five parts: (i) a core of nuclear fuel, (ii) a moderator for slowing down neutrons, (iii) control rods, (iv) coolant or heat exchanger for removing heat in the core, and (v) radiation shielding.

Nuclear fuel is a material that can be fissioned by thermal neutrons. It can be either one or all of the following isotopes: uranium-233, uranium-235 and plutonium-239. We shall see that when natural

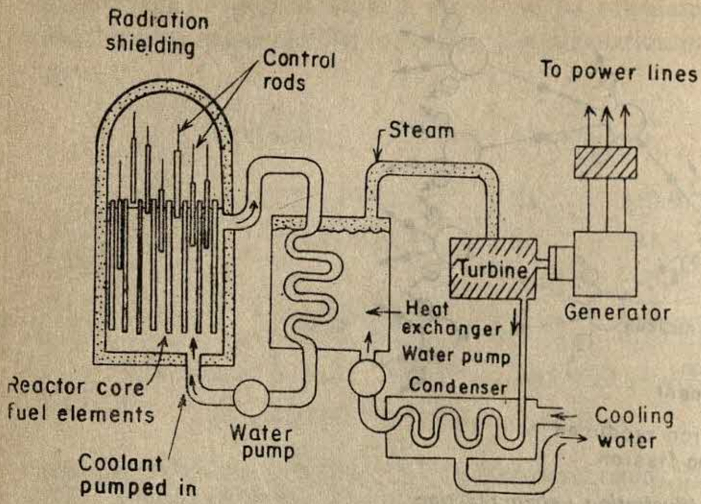


Fig. 1315 *Schematic diagram of a nuclear reactor*

uranium is used, plutonium is produced in the nuclear reactor. Usually, the fuel is put in different aluminium cans in cylindrical rods placed some distance apart. The fuel cans are separated by the moderator. As mentioned earlier, the moderator is used to slow down the fast neutrons produced in the fission process when thermal neutrons strike the nuclear fuel. The fast neutrons have many collisions with the material and come out with thermal energies to strike another fuel can. The material of moderator: (i) should be light, and (ii) should not absorb neutrons. Usually, graphite and heavy water (water containing deuterium instead of hydrogen) are used as moderators.

Sometimes the chain reaction, once started, can liberate an enormous amount of energy and can go out of hand and this can even blow up the reactor. To avoid such an accident, and to regulate the power level of the reactor, control rods are used. These control rods can be inserted into or drawn out of the reactor fuel core and consist of a material that absorbs neutrons, e.g. cadmium, boron or hafnium. Usually, cadmium control rods are used. If these rods are drawn out, the activity of neutrons increases and if they are inserted into the fuel core, the activity of neutrons decreases because they are absorbed by the rods.

The coolant, or heat exchanger, is used to cool the fuel rods and the moderator, and is capable of carrying away large amounts of heat generated in the fission process. If the moderator, fuel rods, etc., are not cooled, the heat generated can melt them. The heat carried by the coolant produces steam that can run a turbine, which in turn can run an electric generator as described in Chapter 7. The last part of the nuclear reactor is the shielding device. Since the neutrons and the fragments in a reactor undergo radioactive decay and produce radiations which are harmful to life, there must be some shielding device to absorb the radiations. For this purpose a concrete wall which is a few feet thick is used.

The nuclear reactor gives a tremendous amount of energy and can develop huge amounts of electric and thermal power. In India, a number of such reactors have been installed. The first atomic reactor at Tarapur was similar to the one described above. The Tarapur reactor produces about 400 MW of electric power.

As mentioned earlier, naturally occurring uranium has only 0.7% of $^{235}_{92}U$ and the rest is $^{238}_{92}U$, which is not fissionable. But it is found that $^{238}_{92}U$ can be made fissionable. This is because of the fact that when fast neutrons strike uranium fuel in a reactor, $^{238}_{92}U$ (present along with $^{235}_{92}U$) absorbs a neutron and becomes $^{239}_{92}U$ which is radioactive and undergoes β-emission twice to get converted into neptunium and plutonium according to the reactions

$$^{238}_{92}U + {}^{1}_{0}n \longrightarrow {}^{239}_{92}U \xrightarrow{-\beta} {}^{239}_{93}Np \xrightarrow{-\beta} {}^{239}_{94}Pu$$

$^{239}_{94}Pu$ itself is fissionable. $^{239}_{94}Pu$ also undergoes α-emission (with a long half life of 24,000 years) to produce $^{235}_{92}U$. Therefore, using an unfissionable material one produces a fissionable material, i.e. additional fuel. $^{239}_{94}Pu$ undergoes fission even with fast neutrons so that there is no need of a moderator in such a reactor. Reactors of this kind are known as *fast breeder reactors* which use thorium (thorium produces fissionable $^{233}_{92}U$) or natural uranium as fuel elements. Such reactors are operated in India also.

13.11 NUCLEAR FUSION

We have seen that if there is a decrease in mass in a nuclear reaction, energy is released and in case of fission this energy can be utilized for many practical purposes. From the binding energy per nucleon v. mass number curve (Fig. 13.11), it was concluded that the binding energy per nucleon for very light and heavy nuclei is small, i.e. these nuclei are not stable. Further, it can be seen that the energy released per nucleon when a heavy nucleus breaks up (~0.8 MeV per nucleon) is less than the energy released when two light nuclei fuse to form a heavier one; e.g. the energy per nucleon in a deuteron is about 1 MeV and it is about 7 MeV in He, so that when two deuterons combine to form a helium nucleus the energy released is about 6 MeV per nucleon, which is quite large as compared to 0.8 MeV in case of heavier nuclei. Therefore it was realized that more energy can be obtained when very light nuclei which lie on the steep side of the curve (Fig. 13.11), are joined or fused to form a heavier one. This process of fusing light nuclei into a heavier one is called *nuclear fusion* and is the basis of the hydrogen bomb which was first exploded in 1952. It is well known that the explosion in a hydrogen bomb is much more violent than in an atom bomb. An idea of the reaction energy can be had from the nuclear reaction that takes place. If four hydrogen atoms fuse to form a helium atom, the fusion reaction can be written as

$$4\,{}^{1}_{1}\mathrm{H} \rightarrow {}^{4}_{2}\mathrm{He} + 2\beta^{+} + Q$$

Using the values of the rest masses of different particles, Q is about 25.7 MeV. Though this is small compared to the energy in a fission process, the energy released per unit mass in fusion is much larger than in fission.

Another important reaction which usually takes place in a hydrogen bomb is the fusion of deuteron (${}^{2}_{1}\mathrm{H}$) and tritium (${}^{3}_{1}\mathrm{H}$) according to the equation

$${}^{3}_{1}\mathrm{H} + {}^{2}_{1}\mathrm{H} \rightarrow {}^{4}_{2}\mathrm{He} + 17.6\ \mathrm{MeV}$$

There is, however, an important difference between fission and fusion as regards the method of starting and maintaining the reaction. It is very easy to start a fission reaction because any thermal neutron (even a stray neutron) can start a chain reaction. But in case of fusion, the light nuclei must have very high speeds in order to overcome the large Coulomb repulsion at very small distances and these speeds must be maintained. To obtain such speeds the material must be heated to very high temperatures and it is found that temperatures of the order of 10^7 °C, (this is the temperature of the interior of the sun), are needed for this purpose. That is why fusion is also called a thermonuclear reaction. In an uncontrolled fusion reaction (as in hydrogen bomb) such high temperatures are momentarily obtained in an atomic explosion of uranium or plutonium. This explosion takes place within the mass of the fusile material.

The above reaction can be controlled and the energy produced can be utilized for useful purposes (just like a nuclear reactor) if it is carried out in the laboratory. For this purpose the energy must be supplied to heat the substance to such high temperatures ($\sim 10^7$°C).

The next problem is the confinement of the substance at these high temperatures. A lot of research in this direction is being done. One of the ways to attain this is to produce a fully ionized gas (called plasma) out of light materials. This plasma can be heated electrically by passing very high currents ($\sim 10^6$ A). The plasma can be confined in what is known as a magnetic bottle. The details of this are beyond the scope of this book.

13.12 MOLECULES

A molecule is made up of atoms; it may consist of atoms of the same element, such as H_2, O_2, N_2, etc. or atoms of different elements, such as NaCl, H_2O, CO_2, etc. A molecule is said to be stable if energy is required to break it into its constituent atoms. The least energy needed to break a molecule into its constituent atoms is called its *dissociation energy* (E).

atoms are held together in a molecule by forces which are essentially electrical in nature. A proper understanding of these forces in related to the structure of the atom itself. In class XI (Chapter 9) you have learnt that the magnitude of the interatomic force depends upon the separation R between the atoms. A certain separation $R = R_0$ corresponds to stable equilibrium.

Bonding in Molecules

Atoms are held together in a molecule by the so-called *bonds* between them. These bonds are due to electrostatic forces between the positive charge of the nucleus and the negative charge of the electrons. As a rule, bonds form if energy is reduced when atoms come near each other. Bonds can be broadly classified into the following four types:

1. *Ionic Bond.* All bonds form because the electrons of the atoms in a molecule come near the nuclei. In some molecules, the electrons tend to be nearer one nucleus than the other nucleus. When this happens the atoms become ions with opposite charges and hence exert on each other a Coulomb force of attraction. It is this force which holds the atoms together. Such a bond is called the *ionic bond*.

The alkali halides are the best examples of the ionic bond. Consider, for example, a sodium chloride molecule; it is diatomic, consisting of a sodium atom and a chlorine atom. When the two atoms are brought closer together, the weekly bound electron from the sodium atom shifts to the chlorine atom. In this process an ion pair is formed—a positive sodium ion and a negative chlorine ion. They are attracted towards each other due to Coulomb attraction and they settle down in an equilibrium state at a separation R_0, called the *equilibrium bond length*. The equilibrium state is the lowest energy state of the system. The repulsion between the electrons in the ions and the repulsion between the nuclei prevents the ions from coming closer to each other than the bond length R_0 [Fig. 13.16(a)].

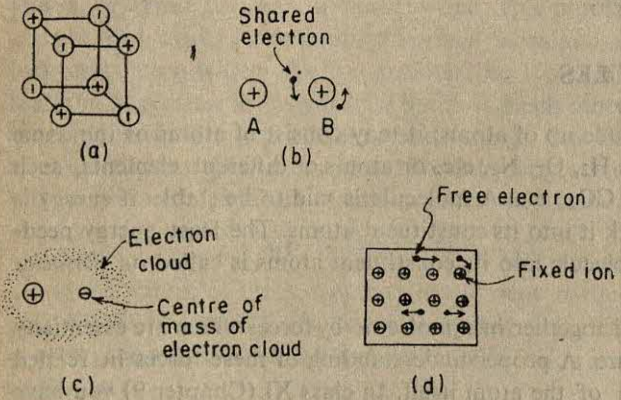


Fig. 13.16 *Diagrammatic illustration of the types of bonds in molecules: (a) ionic bond, (b) covalent bond, (c) van der Waals bond and (d) metallic bond*

We can estimate the dissociation energy E_d of a halide molecule by using the values of the ionization potential, E_i of the alkali atom, electron affinity, E_a of the halogen atom and their equilibrium separation R_0. The ionization potential of an atom is the minimum energy required to remove an electron from it. The electron affinity is the amount of energy released when an atom acquires an extra electron.

Let us again consider the example of a sodium chloride molecule. The ionization potential of sodium atom is $E_i = 5.18$ eV and the electron affinity of chlorine atom is $E_a = -3.62$ eV. From spectroscopic measurements the internuclear separation between Na^+ and Cl^- ions is estimated to be $R_0 = 2.36$ Å $= 2.36 \times 10^{-10}$ m. The dissociation energy E_d of NaCl can be estimated by the following three steps:

(i) First assume that 5.18 eV of energy (which is the ionization potential of sodium) has been spent in ionizing the Na atom, i,e. in releasing an electron from it.

(ii) The released electron is captured by a chlorine atom, liberating 3.62 eV of energy (which is the electron affinity of Cl). Hence the net amount of energy needed to form a pair of Na^+ and Cl^- ions from a pair of Na and Cl atoms is $5.18 - 3.62 = 1.56$ eV.

(iii) The Na^+ and Cl^- ions are attracted towards each other until they achieve equilibrium at a separation $R_0 = 2.36 \times 10^{-10}$ m, so that their Coulomb potential energy is

$$E_c = -\frac{e^2}{4\pi\epsilon_0 R_0} = -\frac{(1.6\times10^{-19})^2}{4\pi\times(8.85\times10^{-12})\times(2.36\times10^{-10})}$$
$$= -9.75\times10^{-19}\text{ J} = -6.09\text{ eV}$$

Therefore energy liberated when Na and Cl atoms form a sodium chloride molecule $= -(-6.09 + 1.56) = 4.53$ eV. The experimentally measured value of the dissociation energy of NaCl molecule is 4.22 eV. Our estimate is 0.31 eV higher than the measured value. This small difference is due to the repulsion between the electrons of Na^+ and Cl^- and between their nuclei and also due to the fact that the ions are not point charges which implies that the value of E_c calculated above is approximate.

2. *Covalent bond* In a covalent bond, the electrons responsible for binding two atoms together are shared between the two nuclei. The best example of the covalent bond is the hydrogen molecule H_2. Detailed calculations show that in the minimum potential energy state, the internuclear separation R_0 between the two hydrogen atoms is 0.74 Å. This stable or equilibrium arrangement is called the hydrogen molecule. The radius of a hydrogen atom is 1.2 Å. Since the equilibrium bond length R_0 is smaller than the atomic radius, the two atoms in a hydrogen molecule cannot be treated as independent and the two electrons are *shared* between them. The electron in one atom, say A, wanders to the other atom B. Thus at any instant, the atom A from which the electron has departed is left with a positive charge and the atom B to which the electron moves has a negative charge. The resulting Coulomb attraction between the two opposite charges holds the two atoms together [Fig. 13.16(b)]. The same result is obtained when the electron moves from B to A at another instant. *The bond between atoms due to shared electrons is called a covalent bond.*

In covalent bonding, a reduction in energy occurs. In the case of a hydrogen molecule, the reduction in energy is about 4.5 eV. This energy has to be supplied from outside in order to dissociate a hydrogen molecule. The reduction in energy is due to special quantum mechanical effects related with the spin angular momentum of electrons which cannot be described in terms of the simple classical theory.

3. *Van der Waals Bond* Atoms of inert gases have closed-shell elec-

tronic structures. Therefore, the bonding between the atoms of inert gases is neither ionic nor covalent. The attraction between neutral atoms can be explained by assuming that the shape of the atom undergoes a fluctuating distortion due to its interaction with other atoms. At any instant, more electrons may appear on one side of the nucleus than on the other. Thus the centre of mass of the electron cloud (i.e. the negative charge) is slightly displaced from the positive charge of the nucleus [Fig. 13.16(c)]. The two charges form an electric dipole. A dipole will attract electrons from the neighbouring atoms, thus inducing other dipoles. The interactions between induced dipoles is responsible for their bonding. The dipole interaction energy varies as $1/R^6$ (where R is the internuclear separation) and is quite small. The forces between induced dipoles are weak; they are called Van der Waals forces in honour of the Dutch physicist C.F. Van der Waals who had predicted similar attractive forces between molecules of gases.

4. *Metallic bond* The three types of chemical bonding described above cannot explain the existence of metals. Metals have free or conduction electrons (electrons outside the closed shell of their atoms) which are free to move around in the metal. One or more electrons in the outermost shells of an atom may leave the atom and occupy the orbit of another atom. These free electrons wander through the structure of the metal which consists of free positive ions [Fig. 13.16(d)]. The electrons are not attached to any particular atom as in the case of a covalent bond.

It may be remarked that the ionic bond is "*non-saturated*" the force of attraction between two ions is not influenced by the surrounding ions. But covalent bonds can be *saturated*. This explains why the atoms in metals are closely packed.

Rotational and Vibrational Spectra

We have seen that the spectrum of an element such as hydrogen is due to the emission of electromagnetic radiations when the excited electrons fall from higher to lower energy states. These electrons are bound to the nucleus by the attractive Coulomb force. Just as the motion of an electron is restricted to orbits which correspond to quantum mechanical stationary states, the rotational and vibrational motions of molecules are also restricted to configurations which correspond to stationary states. Therefore, a molecule has discrete energy levels.

When a molecule is excited from its ground state to higher stationary states (by supplying energy), the excited molecule can make transitions between the stationary states and finally reaches its ground state. In these transitions, photons of different discrete energies (and hence frequencies) are emitted. This is the origin of the rotational and vibrational spectra of molecules which are closely spaced spectral lines. The frequencies corresponding to the various characteristic lines of the spectrum of a molecule are obtained by using quantum mechanics.

SUMMARY

The atom is composed of charged particles. The positive charge is concentrated within a very small volume called the nucleus. The negatively charged particles, called electrons, revolve around the nucleus in discrete orbits, the number of electrons in nth orbit being governed by $2n^2$.

Niels Bohr proposed a new theory of the hydrogen atom and successfully explained the observed spectral lines of atomic hydrogen. According to him, transitions from higher to lower energy states are responsible for emission of radiation that occurs in definite and discrete quanta.

The atomic masses of different elements can be determined by first obtaining the atoms as ions and then observing their deflections in electric and magnetic fields.

Isotopes are atoms of the same element having different atomic masses. The neutron is an uncharged particle whose mass is nearly equal to that of the proton. The nucleus consists of protons and neutrons. The neutrons and protons are packed in a very small volume. The density of the nucleus is very high.

Owing to the binding of the nucleons in the nucleus, the mass of a nucleus is less than the sum of the masses of nucleons and the decrease in mass (called mass defect) appears as the binding energy of a nucleus.

The naturally occurring heavy nuclei are unstable and decay with the emission of α-particles, β-particles and γ-rays. α-particles are ${}_2^4\mathrm{He}$ nuclei, β-particles are electrons, γ-rays are neutral and they are electromagnetic radiations. The decay of any radioactive substance is governed by the exponential decay law given by $N = N_0 e^{-\lambda t}$.

The atoms in a molecule are bound to each other by four types of chemical bonds: covalent, ionic, vander Waals and metallic. Molecules can rotate about an axis and vibrate along a line. These two motions are quantized and are responsible for discrete rotational and vibrational spectra of molecules.

EXERCISES

A. Short-Answer Questions

1. What are alpha particles?
2. What is quantization of a physical quantity?
3. What is the radius of the innermost Bohr orbit?
4. What is the ionization energy of a hydrogen atom in electron volt?
5. State, with unit, the value of Planck's constant.
6. State, in MeV, the energy equivalent of 1 amu.
7. What is the number of electrons, protons and neutrons in a neutral atom whose nucleus is represented by the symbol $^{235}_{92}U$.
8. Name the particles emitted from a radioactive nucleus.

B. Long-Answer Questions

1. Describe briefly Rutherford's alpha particle scattering experiment. What important conclusion did he draw of his observations about the model of an atom?
2. What do you understand by the term "quantization"? Obtain the expression for the discrete values of the energy of a particle of mass m moving in a line of length L between two rigid walls.
3. State the postulates of Bohr in the formulation of his theory of the hydrogen atom and hence obtain the expression for: (a) the radius of the nth Bohr orbit, and (b) the energy of the hydrogen atom in the nth stationary state.
4. Explain what is meant by: a normal atom, (b) an excited atom and (c) an ionized atom.
5. What are Lyman, Balmer and Paschen series in the spectrum of the hydrogen atom? What is their origin?
6. How did Chadwick discover the neutron? How did he estimate its mass?
7. What is meant by atomic number and atomic mass number of a nucleus?
8. What are isotopes? Give two examples.
9. The nuclear symbol for an isotope of carbon is $^{14}_{6}C$. What is the number of (a) protons, (b) electrons, (c) neutrons and (d) nucleons in it?
10. What can you say about the nature and range of nuclear force? How is it different from other known forces?
11. What do you understand by the terms radioactivity and radioactive substances? Name two radioactive elements.
12. What is the nature of α, β and γ radiations? How will you establish experimentally what charge these radiations carry.
13. What changes occur in the nucleus of a radioactive element when it emits (a) an alpha particle and (b) a beta particle? Give an example, in each case, in support of your answer.
14. A radioactive nucleus is represented by the symbol $^{A}_{Z}X$ where A is the atomic mass number and Z is the atomic number. How is the new element represented after the emission of: (a) an alpha particle, (b) a beta particle and (c) a gamma ray?
15. The element $^{238}_{92}U$ emits an alpha particle and becomes thorium (Th). What is the atomic number and atomic mass number of thorium? Write down the equation that represents the change.

16. The element $^{24}_{11}\text{Na}$ emits a beta particle and becomes magnesium (Mg). What is the atomic number and atomic mass number of magnesium? Write down the equation that represents this change.
17. An element $^{202}_{84}\text{X}$ emits an alpha particle and then a beta particle. The final element is $^{a}_{b}\text{Y}$. Find the value of a and b.
18. (a) Name the type of particles which may be given off during radioactive disintegration.

    (b) From which part of the atom do these particles come?

    (c) What changes occur in the radioactive atom when it emits one of these particles (consider the effect of the loss of each kind separately)?

19. Atom X (atomic mass number 238 and atomic number 92) is radioactive and becomes atom Y (atomic mass number 234 and atomic number 90) by the emission of a particle.

    (i) What particle was emitted?

    (ii) How many protons and neutrons does this particle contain?

    (iii) Write down the equation that represents this change.

20. Obtain the radioactive decay law and hence find the expression for the half-life of an unstable radioactive sample. What is the meaning of the statement, "the half life of $^{14}_{6}\text{C}$ is 5700 years"?
21. Explain how carbon dating is used to estimate the approximate age of a tree when it was cut.
22. (a) Explain what you understand by mass defect and binding energy.

    (b) What are the main features of the binding energy per nucleon v. mass number A curve? What is the importance of this curve?

23. (a) What is meant by nuclear force? How does it differ from Coulomb force?

    (b) What do you understand by the statement, "nuclear force is charge-independent"?

24. What is nuclear fission? Describe the fission of $^{235}_{92}\text{U}$ and write down the equation of the fission reaction. What is the source of the energy released in the reaction.
25. What is the chain reaction? Under what conditions is it achieved?
26. In a nuclear reactor, explain the role of: (a) moderator, (b) control rods and (c) delayed neutrons.
27. What is nuclear fusion? Describe the fusion of $^{2}_{1}\text{H}$ and $^{3}_{1}\text{H}$ and write down the equation of the fusion reaction. What is the source of energy released in the reaction? Under what conditions is the fusion achieved?
28. Explain, giving an example of each, the four types of chemical bonding in molecules.

C. Multiple-Choice Questions

Choose the correct answer from the given alternatives:

1. According to Bohr's theory of the hydrogen atom, the radii r_n of stationary electron orbits are related to the principal quantum number n as

    (a) $r_n \propto 1/n^2$ (b) $r_n \propto 1/n$

    (c) $r_n \propto n$ (d) $r_n \propto n^2$

2. The energy of the electron of hydrogen orbiting in a stationary orbit of radius r_n is proportional to

(a) r_n
(b) $1/r_n$
(c) r_n^2
(d) $1/r_n^2$

3. The ionisation potential of the hydrogen atom is 13.6 eV. Its energy in $n = 2$ energy state is

(a) -3.4 eV
(b) -6.8 eV
(c) -13.6 eV
(d) -27.2 eV

4. The innermost orbit of the hydrogen atom has a diameter of 1.06 Å. What is the diameter of the tenth orbit?

(a) 5.3 Å
(b) 10.6 Å
(c) 53 Å
(d) 106 Å

5. Out of the following, which is not emitted by a radioactive substance?

(a) electrons
(b) electromagnetic radiations
(c) helium nuclei with a charge equal to that of two protons
(d) neutrons

6. A radioactive element X has atomic number Z and atomic mass number A. It decays by the emission of an alpha particle and a gamma ray. The new element is

(a) ${}^{A-2}_{Z-1}\text{Y}$
(b) ${}^{A-4}_{Z-2}\text{Y}$
(c) ${}^{A+1}_{Z}\text{Y}$
(d) ${}^{A+4}_{Z+2}\text{Y}$

7. The radioactive decay of uranium into thorium is represented by the equation

$$ {}^{238}_{92}\text{U} \rightarrow {}^{234}_{90}\text{Th} + x $$

What is x?

(a) an electron
(b) a proton
(c) an alpha particle
(d) a neutron

8. A carbon nucleus emits a particle x and changes into nitrogen according to the equation

$$ {}^{14}_{6}\text{C} \rightarrow {}^{14}_{7}\text{N} + x $$

What is x?

(a) an electron
(b) a proton
(c) an alpha particle
(d) a photon

9. The radioactive decay of an element X to elements Y and K is represented by the equation

$$ {}^{A}_{Z}\text{X} \longrightarrow {}_{Z+1}{}^{A}\text{Y} \longrightarrow {}^{A-4}_{Z-1}\text{K} \longrightarrow {}^{A-4}_{Z-1}\text{K} $$

The sequence of the emitted radiations is

(a) α, β, γ (b) β, α, γ
(c) γ, α, β (d) β, γ, α

10. A rate-meter measures the number of disintegrations per second from a radioactive source. It gives a count of 320 counts per second. Ninety minutes later, it gives 40 counts per second. What is the half-life of the source?

(a) 30 minutes (b) 45 minutes
(c) 60 minutes (d) 75 minutes

11. Two radioactive sources A and B initially contain an equal number of radioactive atoms. Source A has a half-life of 2 h. At the end of 2 h, the ratio of the rate of disintegration of A to that of B is

(a) 1 : 2 (b) 2 : 1
(c) 1 : 1 (d) 1 : 4

12. Graphite and heavy water are two common moderators used in a nuclear reactor. The function of the moderator is

(a) to slow down the neutrons to thermal energies
(b) to absorb the neutrons and stop the chain reaction
(c) to cool the reactor
(d) to control the energy released in the reactor

13. Cadmium rods are used in a nuclear reactor for

(a) slowing down fast neutrons
(b) speeding up slow neutrons
(c) absorbing neutrons
(d) regulating the power level of the reactor

14. A uranium nucleus (atomic number 92, mass number 238) emits an alpha particle and the resultant nucleus emits a β-particle. The atomic and mass numbers respectively of the final nucleus are

(a) 90,240 (b) 90,236
(c) 91,234 (d) 92,232

15. The chemical behaviour of an atom depends upon

(a) the number of electrons orbiting around its nucleus
(b) the number of protons in its nucleus
(c) the number of neutrons it its nucleus
(d) the number of nucleons in its nucleus

D. Numerical Problems

1. If the innermost orbit of hydrogen atom has a diameter of 1.06 Å, what is the diameter of the 10th orbit?
2. Calculate the energy of the hydrogen atom in the states $n = 4$ and $n = 2$. Compute the frequency and wavelength of the emitted radiation in a transition from $n = 4$ to $n = 2$ state. Is this radiation visible?

3. A hydrogen atom makes a transition from $n = 3$ to $n = 2$ state. What is the wavelength of the emitted photon? Is this photon of visible light?
4. In a mass spectrometer, singly ionized chlorine ions pass through its velocity selector which has an electric field of 50 kV m^{-1} and a magnetic field of 0.5 Wb m^{-2} at right angles to each other and emerges undeflected. Calculate the radius of curvature of its path in the magnetic field if the mass of chlorine ion is 34.980 u.
5. The fusion reaction

$$^{2}_{1}H + ^{2}_{1}H \rightarrow ^{4}_{2}He + Q$$

is proposed to be used for the production of industrial power. Assuming the efficiency of the process to be 30%, find how many grams of deuterium will be consumed in a day for an output power of 50,000 kW. Given mass of $^{2}_{1}H = 2.01478$ u and of $^{4}_{2}He = 4.00388$ u.

6. The binding energy per nucleon of ^{16}O is 7.97 MeV and that of ^{17}O is 7.75 MeV. Calculate the energy required to remove a neutron from ^{17}O.
7. An accident occurs in a laboratory in which a large amount of radioactive material, with a known half-life of 20 days, becomes embedded in the floor and walls, etc. Tests show that the level of radiation is 32 times the permissible limit. After how many days can be laboratory be safely occupied?
8. The decay rate of $^{200}_{79}Au$ of mass 20×10^{-9} kg is 60 Ci. Calculate the half-life of the sample.
9. 10 mg of carbon from living matter has the decay rate of 200 disintegrations per minute due to a small proportion of the radioactive ^{14}C. A piece of ancient wood of mass 10 mg is found to have a decay rate of 50 disintegrations per minute. Find the age of the wood, assuming that the content of ^{14}C in the atmosphere has remained unchanged. The half-life of ^{14}C is 5700 years.
10. The half-life of radium is 1500 years. In how many years will 1 g of pure radium: (a) lose 1 mg, and (b) be reduced to 1 centigram?
11. Find the volume at NTP of radon $^{222}_{86}Rn$ whose half life is 3.825 days which will have an activity of 0.1 curie.
12. Radioactive isotopes ^{32}P and ^{33}P are mixed in the ratio 2 : 1 by the number of atoms. The activity is 2 microcuries. Find the activity after 30 days. Given half life of ^{32}P = 14 days and that of ^{33}P = 25 days.
13. Calculate the number of alpha decays that occur in a one gram sample of ^{232}Th in one year. The disintegration constant of $^{232}Th = 1.58 \times 10^{-18}$ s^{-1}. What is the activity of the given sample?
14. Two consecutive radiations from a long-lived parent have decay constants λ_1 and λ_2. Show that the apparent mean life of the second radiation is

$$\left(\frac{1}{\lambda_1} + \frac{1}{\lambda_2}\right)$$

15. The fission of $^{235}_{92}U$ produces a mean energy of 195 MeV per nucleon. Calculate the mass of $^{235}_{92}U$ that would be required per day to run a generating stations working at 30% overall efficiency and producing 100 MW of power.
16. The fission of $^{235}_{92}U$ produces a mean energy of 185 MeV. If $^{235}_{92}U$ in a reactor is continuously generating 100 MW of power, how long will it take for 1 kg of uranium to be used up?

17. The sun and the stars are believed to be 'getting' their energy from the fusion of four protons to form a helium nucleus and a pair of positrons. Calculate the release of energy per fusion and the rate at which hydrogen must be consumed to produce a kilo-joule of energy per second. Given

$$^{1}_{1}\text{H} = 1.007825 \text{ u (hydrogen atom)}$$

$$^{4}_{2}\text{He} = 4.002603 \text{ u (helium atom)}$$

$$m_{\beta^+} = m_{e^-} = 5.5\times10^{-4} \text{ u (positron)}$$

The energy carried away by the accompanying neutrinos is assumed to be negligible.

18. Calculate the force between two fission fragments of equal masses and sizes that are produced in the fission of ^{239}Pu (by a thermal neutron) in which 4 neutrons are emitted.

14

SOLIDS AND SEMICONDUCTOR DEVICES

PART A—SOLIDS

14.1 INTRODUCTION

We know that the atoms in a solid have almost fixed locations. It is this lack of mobility of the atoms within a solid which endows it with a definite shape and size. In our study of solids, we will ignore the structure of atoms (or molecules) constituting the solid and treat them as points. The most remarkable thing about a solid is its periodicity. By this we simply mean that a basic pattern is repeated endlessly to form the solid. A crystalline solid is thus characterized by the presence of long-range order in its structure. Thus, a wall-paper spread out over a wall would be the analogue of a two-dimensional solid. Sometimes, the external features reflect the inner symmetry of the solid. For example, the binding strength of the atoms is usually greater in certain directions than in others. Hence, there are certain planes through the crystal called cleavage planes, where it is more easily broken compared to others. In certain cases, the inner structure manifests itself on the surface in some form. A quartz crystal, for example, does not have sides of equal length, but the angles between the faces are exactly 120°. Hence, it reflects the inner hexagonal symmetry. However, it was only after the development of the powerful technique of x-ray diffraction that it became possible to determine in detail the geometry of the internal patterns of crystals.

Solids which exhibit only short-range order are called *amorphous* or *glassy solids*. In class XI you have studied these types of solids. In this chapter we will deal only with crystalline solids and learn about them in a little more detail.

14.2 CRYSTAL STRUCTURE

A crystalline solid is made up of a three-dimensional array of atoms.

It can be built up by stacking together a large number of identical structural units. Hence, in order to describe the structure of a crystal, it is not necessary to specify the positions of all the atoms in it.

Let us take the simplest (hypothetical) example of a structure in a plane (i.e. in two dimensions) as shown in Fig. 14.1(a). Each structural unit consists of two atoms, represented by an open circle and a filled (blackened) circle. This basic unit is repeated periodically.

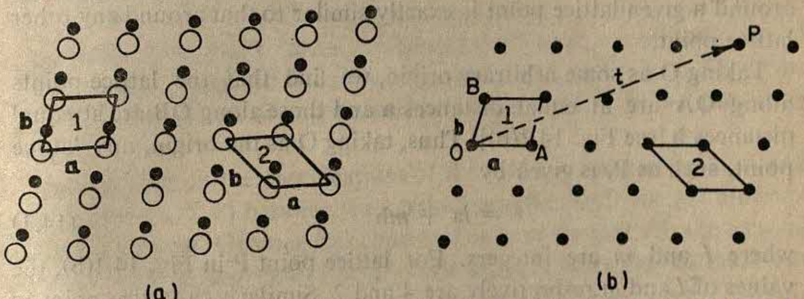


Fig. 14.1 *(a) Arrangement of two-atom groups in a two-dimensional crystal, (b) its crystal lattice and two possible unit cells; vectors* **a** *and* **b** *define a unit cell*

The Unit Cell

The fact that all the atoms in a crystal are arranged in a regular and orderly manner makes it possible to describe the structure of a crystal by considering only a small group of atoms and imagining them to be located in a cell called the *unit cell.* It is the smallest structural unit which repeats itself to form the entire crystal. Thus *a unit is the smallest structural unit of the arrangement of atoms in a crystal from which the entire crystal can be built up by translating the unit parallel to itself in the direction of its edges by distances equal to the lengths of the edges.* A unit cell can be chosen in many different ways. Figure 14.1(b) shows two possible ways of choosing a unit cell. It is preferable to choose a unit that reflects the symmetry of the crystal. Thus the knowledge of the location of atoms in a unit cell and the manner in which the cell is repeated gives the complete description of the structure of the crystal. The lengths of the edges (or the axial distances **a** and **b**) of the unit cell are called the *translation vectors.*

The Crystal Lattice

The study of the geometrical arrangements of atoms in a crystal is considerably simplified if we replace all the repeating units (atoms or molecules) by mere points. The array of points replacing all the groups of atoms is called a *crystal lattice* or *Bravais lattice* (after the French physicist Auguste Bravais, who introduced the concept of lattice in 1848). Thus *a crystal lattice is a large assembly of points, each point representing the repeating group of atoms in the crystal.* Figure 14.1(b) shows a two-dimensional lattice. Each point in the lattice is called the *lattice point* or *lattice site.* Because of the regularity and orderliness which characterize a crystal structure, the arrangement of points around a given lattice point is exactly similar to that around any other lattice point.

Taking O as some arbitrary origin, we find that the lattice points along OA are at equal distances **a** and those along OB are at equal distances **b** [see Fig. 14.1(b)]. Thus, taking O as the origin, any lattice point, such as P, is given by

$$\mathbf{t} = l\mathbf{a} + m\mathbf{b} \tag{14.1}$$

where l and m are integers. For lattice point P in Fig. 14.1(b), the values of l and m respectively are 4 and 2. Similarly, a lattice point in a three-dimensional lattice is, in general, given by

$$\mathbf{t} = l\mathbf{a} + m\mathbf{b} + n\mathbf{c} \tag{14.2}$$

where n (like l and m) is an integer and **a**, **b** and **c** are the translation vectors.

Figure 14.2 shows how the three-dimensional crystal can be obtained by identically and periodically repeating the same unit cell many times in three different directions in space.

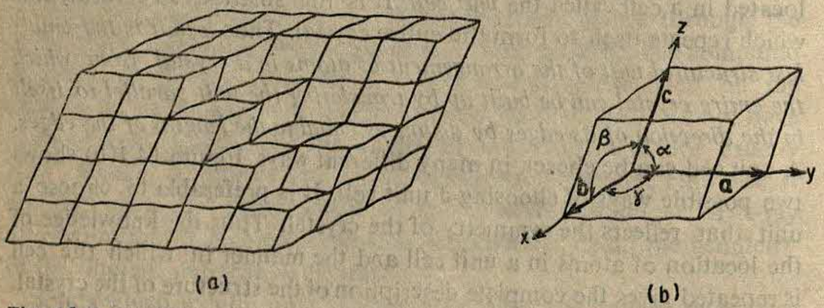


Fig. 14.2 *(a) Arrangement of unit cells in a three-dimensional crystal (b) A unit cell described by three lattice constants* **a**, **b** *and* **c** *and angles α, β and γ between them*

Figure 14.2(b) shows a unit cell in three dimensions. It is a parallelepiped defined by the three axes **a**, **b** and **c**. The entire lattice can be constructed from this fundamental building block by repeated translations along its edges. The directions of the lattice constants **a**, **b** and **c** are called the *crystallographic axes.*

Different Systems of Crystal Structure

In the determination of crystal structure, the sizes and shapes of the crystals are not of much significance. What are really important are the angular orientations of the different faces of the crystal. These orientations are represented by taking a point in a crystal and drawing from it perpendiculars to different faces to get a radiating set of three axes. These are the crystallographic axes, drawn parallel to the lines of intersection of any three faces which do not lie in the same plane.

Depending on the specifications of the lattice constants (**a**, **b**, **c**) and the angles (α, β, γ) between the crystallographic axes, we get different types of lattices. Geometrical considerations show that all the crystal lattices can be grouped in seven basic systems according to the seven different types of unit cells we can choose for them. These systems can be distinguished from one another by the angles α, β and γ and the lattice constants **a**, **b** and **c**. The seven basic crystal systems are triclinic, monoclinic, orthorhombic, tetragonal, cubic, hexagonal and rhombohedral. They are listed in Table 14.1.

Table 14.1 Basic crystal systems

System	Features	Examples
1. Triclinic	$\mathbf{a} \neq \mathbf{b} \neq \mathbf{c};\ \alpha \neq \beta \neq \gamma$	$K_2Cr_2O_7$
2. Monoclinic	$\mathbf{a} \neq \mathbf{b} \neq \mathbf{c};\ \alpha = \beta = 90°,\ \gamma \neq 90°$	$KClO_3$
3. Orthorhombic	$\mathbf{a} \neq \mathbf{b} \neq \mathbf{c};\ \alpha = \beta = \gamma = 90°$	$HgCl_2$
4. Tetragonal	$\mathbf{a} = \mathbf{b} \neq \mathbf{c};\ \alpha = \beta = \gamma = 90°$	White tin
5. Cubic	$\mathbf{a} = \mathbf{b} = \mathbf{c};\ \alpha = \beta = \gamma = 90°$	Li, Ni, NaCl, Ag
6. Hexagonal	$\mathbf{a} = \mathbf{b} \neq \mathbf{c};\ \alpha = \beta = 90°,\ \gamma = 120°$	Zn, quartz
7. Rhombohedral	$\mathbf{a} = \mathbf{b} = \mathbf{c};\ \alpha = \beta = \gamma \neq 90°$	Calcite

The structures of these seven systems are illustrated in Fig. 14.3.

It further follows from geometrical considerations that there are only 14 possible ways in which similar points can be arranged in a regular manner in a three-dimensional space. Thus there are only 14 possible lattices for single atoms. Each of the seven crystalline systems mentioned can be associated with a definite number of space lattices. For example, the cubic crystalline system has the following three possible types of space lattices, depending upon the shape of the unit cell.

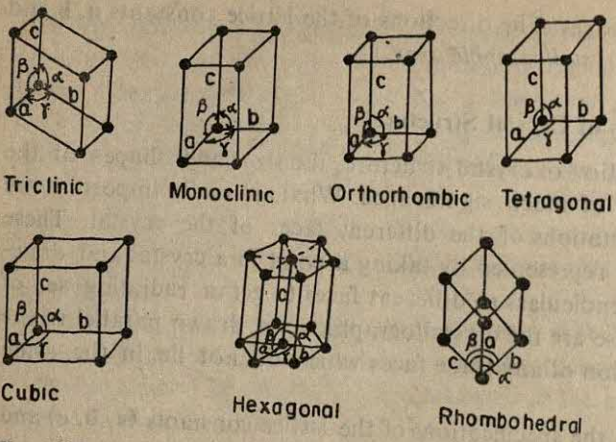


Fig. 14.3 *Unit cells of the seven lattice systems*

1. *Simple Cubic Lattice (sc)* In this, the atoms lie only at the corners of each unit cube [Fig. 14.4(a)].

2. *The face-centred cubic lattice (fcc)* In this, there are atoms lying at the corners and also at the centre of each face [Fig. 14.4(b)].

3. *The body-centred cubic lattice (bcc)* In this, there is also an atom lying at the centre of the unit cell, besides the atoms lying at the corners [Fig. 14.4(c)].

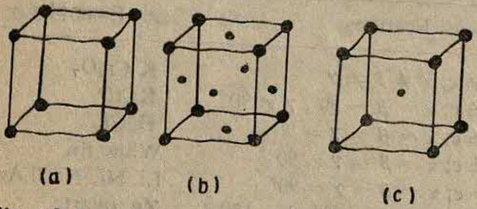


Fig. 14.4 *(a) Simple cubic lattice*
(b) Face-centered cubic lattice
(c) Body-centered cubic lattice

A detailed discussion of the various lattice types would carry us far away from the boundaries of an elementary treatment. However, we can make some general remarks. The lattice with the least symmetry is the triclinic lattice. It corresponds to the oblique lattice in two dimensions. The alkali metals (Li, Na, K, Rb and Cs) have body-centred cubic lattices. The monovalent noble metals (Cu, Ag, Au) form face-centred cubic lattices. For ionic crystals, the two common types of structures are the sodium chloride (NaCl) structure and the caesium chloride (CsCl) structure. Another commonly encountered structure is the diamond structure which is exhibited by covalent semiconductors (Si and Ge). The inert gas crystals (molecular crystals) are of the face-centred cubic type.

Three Kinds of Crystals

1. *Single Crystals* Crystals in which the periodicity of the pattern extends throughout the piece of a crystal are called *single crystals*. The size of a single crystal can be as big as a few centimetres or as small as a fraction of a centimetre. Such small crystals, whose features cannot be seen with the naked eye, are called *monocrystals*. The dimension of a typical unit is of the order of 10^{-8} cm (about 1Å). The size of a monocrystal having as many as 10^4 unit cells is only about 10^{-4} cm.

Like all crystalline solids, single crystals are anisotropic, i.e. their physical properties, such as mechanical strength, thermal and electrical conductivity, are different in different directions. Amorphous solids, on the other hand, are *isotropic*, i.e. their physical properties do not depend on the direction.

2. *Polycrystals* A polycrystal is an aggregate of monocrystals joined together. If a sugar grain is examined under a microscope, we can see small crystals with well developed faces joined together.

Metals are polycrystalline. When a molten metal solidifies, the monocrystals grow with their axes oriented in random directions. Consequently, a polycrystalline metal exhibits isotropic properties, though the monocrystals constituting them are anisotropic.

Ceramics are an important class of polycrystalline solids. Porcelain cups, plates, tiles, etc. are common examples of ceramics. Ceramics are made of inorganic oxides such as sand (SiO_2). They are easy to melt. To get a big solid block of such a material, it is crushed to form a power which is then mixed with binder and pressed to get desired shapes. It is then heated to a high temperature and cooled. The binder evaporates and we get a solid piece called a *ceramic*. By a special technique, ceramics can be made anisotropic. They find many applications in gas lighters and telephone receivers. Ceramics made by suitably combining materials like Y_2O_3, BaO and CuO_2 show superconductivity which has promising uses, especially in power transmission.

3. *Liquid Crystals* When some organic crystalline solids are heated, they do not directly go to the liquid phase. In a certain range of temperature between the solid and the liquid phases, they become

fluids, like liquids, but their molecules remain oriented in particular directions which implies that they retain their highly anisotropic properties. Such fluids which have a long-range orientational order are called *liquid crystals.*

In the liquid crystal phase, some materials (such as cholesterol esters) change their colour with slight changes in temperature. This change in colour with temperature is used to measure temperature in liquid crystal thermometers. Some liquid crystals (especially alkali cyanobiphenyl) change the plane of polarization of light incident on them. They are used as liquid crystal displays (LCD) in electronic watches, clocks, calculators, etc.

14.3 ENERGY BANDS IN SOLIDS

The free electron model of a metal explains, in a qualitative manner, some of the observed properties of metals. However, it does not explain the enormous range of resistivities exhibited by solids. Why are some materials good conductors of electricity, while others (insulators) are poor conductors? Why should some materials, called semiconductors, have conducting properties somewhere in between conductors and insulators? The band theory of solids is remarkably successful in answering these questions.

We have seen that electrons are arranged in various shells around the atomic nucleus. The valence electrons occupy the outermost shell in an atom. The core electrons are tightly bound to the nucleus and are not easily disturbed by external influences. When energy is supplied to an atom (thermal, electrical, or any other type), it is the valence shell which first receives this energy. To discuss electrical conduction in a solid we must study how the electrons respond to an applied electric field. Clearly, the core electrons are not going to play any major role in this.

An atom tends to have a filled valence shell. Thus, alkali metals, which have just one valence electron, are the best conductors. If the valence shell is more than half full, the atom tends to fill its shell. Therefore, elements with five or more valence electrons (like P, O, S, etc.) are, generally, insulators. Similarly, elements with three or less valence electrons are good conductors. An interesting case is that of the elements Si and Ge which have four valence electrons. Those elements which are neither good conductors nor good insulators are termed semiconductors. Inert gases have filled valence shells and hence behave as very good insulators. For greater precision in these considerations, we now turn to a discussion of the energy bands in a solid.

We have seen in Chapter 13 that an isolated atom has a discrete set of energy levels that can be occupied by the electrons of the atom. When atoms are brought together to form a solid, the motion of the valence electrons is profoundly disturbed due to the electric fields produced by the neighbouring atoms. As a result, the energy levels get distorted. To illustrate the basic ideas, let us consider a specific example. Sodium ($Z = 11$) has the electronic configuration $1s^2 2s^2 2p^6$ $3s$. Figure 14.5 is a schematic representation of the energy levels of electrons in a sodium atom. When two sodium atoms are brought close together, each of the valence electrons is affected by the field of both the Na^+ ions. As a consequence of this the system of two atoms may have two levels which are close together, as shown in Fig. 14.5(b). If more and more atoms are brought together the result is that each energy level splits into a large number of levels. The $1s$ states are deep within the atom and are well shielded by the outer electrons of their own atom. As a result they are not greatly influenced by the neighbouring atoms. In a solid there are about 10^{23} atoms/cm^3 and each atom contributes a valence electron in the case of sodium (and other monovalent elements). Therefore, each energy level of an isolated atom gets split into an enormous number of levels, closely spaced together. Hence, the net result is that the levels form a continuous distribution in an energy band. This is shown in Fig. 14.5(c)

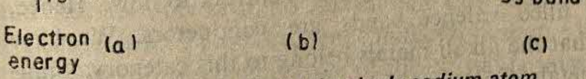


Fig. 14.5 *(a) Energy levels of a single sodium atom*
(b) The 3s level splits into two levels
(c) In a sodium crystal the 3s level spreads into an energy band

The highest energy band occupied by the valence electrons is called the *valence band*. Above this band lies an unfilled band called the *conduction band*.

14.4 CONDUCTORS, INSULATORS AND SEMICONDUCTORS

To participate in the conduction process, an electron must be given sufficient energy so that it can reach the conduction band. A material can be classified as a conductor, insulator, or a semiconductor on the basis of its energy band diagram. The various possibilities which arise in practice are sketched in Fig. 14.6. We now discuss these one by one.

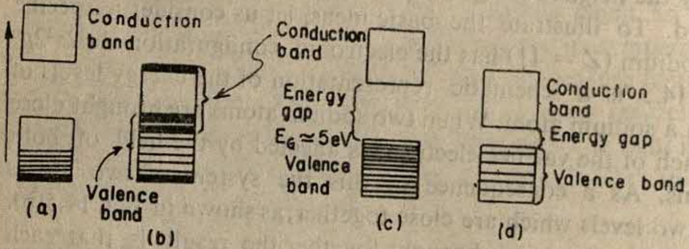


Fig. 14.6 *Energy band diagrams: (a) and (b) conductors, (c) insulator, and (d) semiconductor*

Conductors

In case (a) we find that the valence band is partly filled. This is the situation in the case of sodium and other alkali metals. The reason for this is very simple. A sodium crystal with N atoms will have N possible energy levels within the valence band. Since each energy level can accommodate two electrons with opposite spins, without violating the exclusion principle, the valence band will be full only if it has the maximum allowed quota of $2N$ electrons. Since there is only one valence electron $3s$, the band is only half full. Under the action of an external field, the uppermost electrons in the valence shell can gain kinetic energy and occupy any of the large number of available empty levels within the band. This causes an electric current to flow. Hence, solids with partly filled valence bands are conductors. We have already remarked that the alkali metals belong to this category. Similarly, the noble metals Cu, Ag, and Au have half-filled bands. The elements Al, Ga, In, and Tl have three electrons per atom. These can fill one and a half bands. Thus, these elements are conductors.

In case (b) we have a situation where the valence band and the conduction band overlap. Hence, electrons from the valence shell can easily pass into the conduction band. Materials with such a band structure are, therefore, conductors. Beryllium ($Z = 4$; $1s^2\,2s^2$) and magnesium ($Z = 12$; $1s^2\,2s^2\,2p^6\,3s^2$) belong to this category. Although the valence band is full, the next unoccupied band ($2p$ for Be, and $3p$ for Mg) partly overlaps the valence band. This is possible because an electron in the first excited level differs very little in energy from an electron in the valence shell.

Insulators

The energy band diagram in case (c) shows that the valence band and the conduction band are separated by an energy gap E_G. The region represented by E_G is called the *forbidden band.* An electron in a crystal lattice cannot have an energy which lies in this region. If E_G is large, an applied electric field cannot give enough energy to an electron in the valence band to enable it to enter the conduction band. Hence, the valence electron will not be freed and no current will flow. Thus, materials having a large band gap are insulators. For example, in diamond $E_G \simeq 9$ eV.

Note: It is not possible for us to explain the reason for the existence of these forbidden bands, without introducing quantum-mechanical concepts. If we consider an electron to be a wave propagating in the lattice, the Bragg reflections from the crystal planes can be shown to lead to these gaps.

Semiconductors

Case (d) is similar to (c) except that the band gap is much smaller, typically about 1 eV. Hence, electrons in the valence band may acquire thermal energy to pass on into the conduction band. Such materials can, therefore, carry an electric current but not as easily as a conductor. They are called semiconductors. The best known examples are silicon ($E_G \simeq 1.17$ eV) and germanium ($E_G \simeq 0.74$ eV).

At absolute zero of temperature, the conduction band of semiconductors, is totally empty and all the energy states in the valence band are filled. Due to the absence of electrons in the conduction band near absolute zero, no current can flow under the influence of an electric field. Therefore, semiconductors become insulators at low temperatures. If the temperature of the crystal is raised to say, room temperature, some valence electrons acquire thermal energy greater than the energy gap E_G and move to the conduction band. Hence, even a small electric field can cause a flow of current. Thus a crystal, which was an insulator at low temperatures, becomes slightly conducting at room temperatures. Such crystals are, therefore, called semiconductors. Unlike metallic conductors, the resistance of semiconductors decreases with increase in temperature.

14.5 ELECTRONS AND HOLES IN SEMICONDUCTORS

As mentioned earlier, the electrons in a semiconductor occupy the valence band which is separated from the conduction band by a band gap $E_G \simeq 1$ eV, which is much smaller than the band gap in insulators ($E_G \simeq 9$ eV). This means that the covalent bond in a semiconductor is not very strong. Hence, it is possible, even at room temperature, for a valence electron to break loose from the covalent bond by thermal agitation and enter the conduction band (Fig. 14.7).

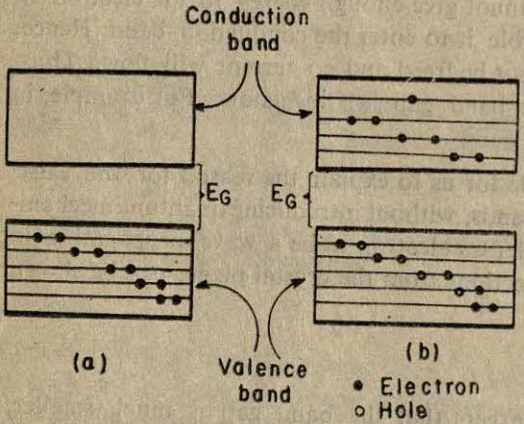


Fig. 14.7 *Energy band diagram of a semiconductor at (a) absolute zero, and (b) room temperature*

As the temperature increases, more valence electrons acquire sufficient energy to become free and move to the conduction band. Therefore the electrical conductivity increases with temperature or resistivity decreases with temperature. As an electron leaves the valence band, it leaves a vacancy behind. This is shown as an open circle in Fig. 14.7(b); the filled circle represents an electron. The vacancy may be regarded as a positive charge equal in magnitude to the charge on an electron and is called a *hole*. It will attract other electrons. Hence, another valence electron may move to fill up this vacancy, leaving a hole behind. In this manner a hole moves through the crystal lattice.

Let us consider the example of a germanium crystal and see how holes are produced. The germanium atom has four valence electrons. The core of each atom (consisting of the nucleus and the core electrons) has four nearest neighbours arranged around it at the corners of a tetrahedron, and it forms covalent bonds by sharing an electron pair with each of them. This is represented schematically in the two-dimensional diagram of Fig. 14.8. The valence electrons occupy filled energy bands. In the case of germanium the energy gap $E_G \simeq 0.75$ eV. Even at ordinary temperatures, the thermal energy of a valence electron is enough to enable it to break loose from the covalent bond and move to the conduction band, leaving behind a vacancy (called hole) in the valence band. This is shown by an open circle in Fig. 14.8. An electron from a neighbouring atom can break away and be attracted by the positive hole (missing electron) thus completing the covalent bond and creating a hole at another place. In our two-dimensional example, we see that an electron from any of the eight neighbouring atoms can come to complete the bond and the hole can move to any of these atoms.

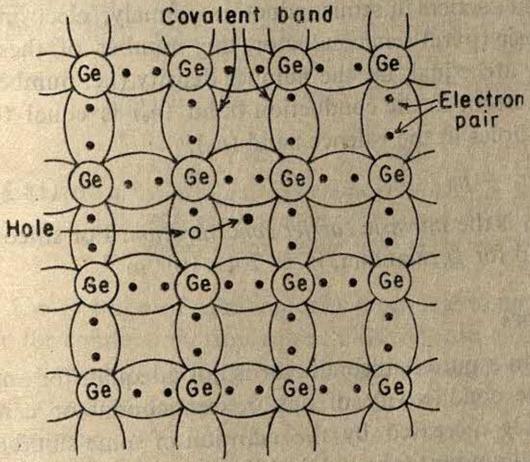


Fig. 14.8 *Breaking of a covalent bond in a germanium (Ge) crystal*

14.6 INTRINSIC AND EXTRINSIC SEMICONDUCTORS

Pure semiconductors are called *intrinsic* semiconductors, e.g. pure germanium and pure silicon. When an electric field **E** is applied across a crystal of a semiconductor, the electrons (being negatively charged) drift in a direction opposite to that of **E** and the holes (which have a positive charge) drift in the direction of **E**. This is shown schematically in Fig. 14.9 in two dimensions.

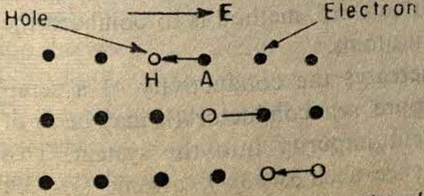


Fig. 14.9 *Current carriers in a semiconductor*

An electron A moves to the left (in a direction opposite to E) to fill a hole at H. This leaves behind a vacancy at A and another electron may move to the left to fill the new hole (shown in the second row, for convenience) and so on. Thus we see that, under the influence of an electric field, the electrons move in one direction (opposite to **E**) and the holes move in the opposite direction (in the direction of **E**). Since the conventional current is opposite to the electron flow, it is clear that the currents due to the flow of electrons and holes add up to give a total current in the direction of the electric field. Thus there are two types of current carriers in semiconductors, namely, electrons and holes. In an intrinsic (pure) semiconductor, the number of these two types of carriers are equal, i.e. the number density (i.e. number per unit volume) of electrons in the conduction band (n_e) is equal to the number density of holes in the valence band (n_h),

$$n_e = n_h = n_i \tag{14.3}$$

This number density n_i is the *intrinsic carrier concentration*. For silicon $n_i \simeq 1.5 \times 10^{16}\ m^{-3}$ and for germanium, $n_i = 2.4 \times 10^{19}\ m^{-3}$.

Extrinsic Semiconductors

The current produced in a pure semiconductor is not adequate for any useful application. The conductivity of a pure semiconductor can, however, be enormously increased by the addition of some suitable impurity in a very small amount (about 1 in 10^6 parts of the semiconductor). Pentavalent and trivalent impurities are suitable. The drastic increase in the conductivity of a semiconductor due to the addition of pentavalent or trivalent impurities gives rise to an enormous range of practical applications of semiconductors. The deliberate addition of a desirable impurity to a semiconductor is called *doping* and the impurity atoms added are called *dopants. A doped semiconductor* (*i.e. a semiconductor with impurity atoms*) *is called an extrinsic semiconductor*.

There are various ways by which doping is achieved in practice. One method is to add the impurities to a molten semiconductor. Another is to heat the semiconductor in the atmosphere containing dopant atoms (or molecules containing dopant atoms) so that the dopants diffuse into the semiconductor. The third method is to bombard the semiconductor with ions of dopant atoms.

Let us now see how doping increases the conductivity of a semiconductor. Suppose we have a pure semiconductor (it may be Si or Ge). Let us introduce a pentavalent impurity into the system. This could be phosphorus ($Z=15$; valence shell $3s^2\ 3p^3$), arsenic ($Z=33$; valence shell $4s^2\,4p^3$), or antimony ($Z=51$; valence shell $5s^2\,5p^3$). If the impurity is added carefully, it will replace one of the ions in the crystal lattice. This is shown in Fig. 14.10(a). The impurity has five valence electrons. Four of these will form covalent bonds, but one excess electron will be left free. Hence, a pentavalent impurity is called a *donor* impurity, since it donates an electron that then takes part in conduction.

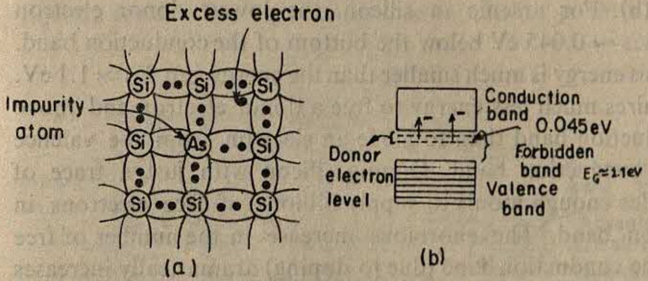


Fig. 14.10 *(a) Addition of arsenic (As) impurity to a silicon (Si) semiconductor (b) Excess electrons introduce donor energy levels close to the conduction band*

The addition of each impurity atom makes an extra electron available for conduction. Since these electrons are not released by thermal agitation, they do not produce holes. Consequently, the current carriers are primarily electrons. Hence, such semiconductors are called *n-type semiconductors* (*n* stands for negative current carriers). Since these additional electrons cannot be accommodated in the valence band, they occupy same discrete energy levels just near the bottom of the conduction band. Hence, electrons can be easily excited from these levels into the conduction band [Fig. 14.10(b)]. Thus, it is much easier to cause electron flow in an *n*-type semiconductor than in pure Ge or Si. This is explained as follows:

The excess electron of the donor atom orbits around the donor nucleus, just as the electron in a hydrogen atom orbits around its nucleus (proton). One major difference is that the electron is not moving in free space (as in a hydrogen atom) but in the crystalline medium of silicon (or germanium). Silicon has a relative permittivity (called its dielectric constant) $\epsilon_r \simeq 10$. We know that the Coulomb force is inversely proportional to ϵ_r (see chapter 1). Hence the Coulomb attraction between the donor electron and the donor nucleus will be 10 times weaker than the electron-proton attraction in a hydrogen atom. We can easily work out Bohr's theory for a hydrogen atom for the case when $\epsilon_r \neq 1$ (as in free space). We will then find that the lowest hydrogen bound state energy is reduced by a factor ϵ_r^2 (which is about

100 for silicon). Primarily because of this reason, the binding energy of the excess electron in arsenic doped silicon is 0.045 eV rather than 13.6 eV (which is the hydrogen atom binding energy). Thus supplying as little energy as 0.045 eV will free the excess electron. This electron then has the lowest possible energy in the empty conduction band. Thus the energy-level diagram of a doped semiconductor is as shown in Fig. 14.10(b). For arsenic in silicon, the lowest donor electron energy level lies $\sim$ 0.045 eV below the bottom of the conduction band. Notice that this energy is much smaller than the energy gap $E_G \simeq 1.1$ eV. Hence it requires much less energy to free a donor electron and move it to the conduction band than to move an electron from the valence band to the conduction band. Doping silicon with just a trace of arsenic provides enough atoms to supply billions of free electrons in the conduction band. The enormous increase in the number of free electrons in the conduction band (due to doping) dramatically increases the conductivity of the extrinsic (doped) semiconductor. Thus, at room temperature, almost all the excess donor electrons are in the conduction band.

The number densities of conduction band electrons (n_e) and the valence band holes (n_h) in a doped semiconductor are very different from those in pure semiconductors (in which $n_e = n_h = n_i$; the intrinsic concentration). From thermodynamical considerations, one can show that

$$n_e n_h = n_i^2 \tag{14.4}$$

In an *n*-type semiconductor, the donor electrons are almost all free at room temperature which means that the donor electrons are the main current carriers and $n_e = n_d$, the donor electron number density. For most common doped semiconductors n_d is much larger than the number of 'intrinsic' conduction band electrons ($n_i = n_h$). Thus

$$n_e = n_d \gg n_h \tag{14.5}$$

Hence, in *n*-type semiconductors, electrons are the *majority carriers* and the holes are the *minority carriers* of current.

When a trivalent impurity is added to Ge or Si, it has a different effect. Some typical trivalent impurities are boron ($Z = 5$; valence shell $2s^2\,2p$), aluminium ($Z = 13$; valence shell $3s^2\,3p$), gallium ($Z = 31$; valence shell $4s^3\,4p$) and indium ($Z = 49$; valence shell $5s^2\,5p$). When an impurity ion replaces a Ge or Si ion, it can only provide three valence electrons for forming covalent bonds. Thus there is a gap left in one of the covalent bonds. This vacancy acts as a hole that tends to accept electrons to fill the gap [Fig. 14.11(a)]. Thus, trivalent impurities are called *acceptor* impurities. In this case the impurity atoms introduce unoccupied discrete energy levels, very close to the top of the valence band. Electrons from the valence band can, therefore, be easily excited by thermal agitation to one of these acceptor hole levels, leaving holes behind them in the valence band [Fig. 14.11(b)]. These holes can then act as carriers of current. Hence, such semiconductors are called *p*-type semiconductors (*p* for positive carriers).

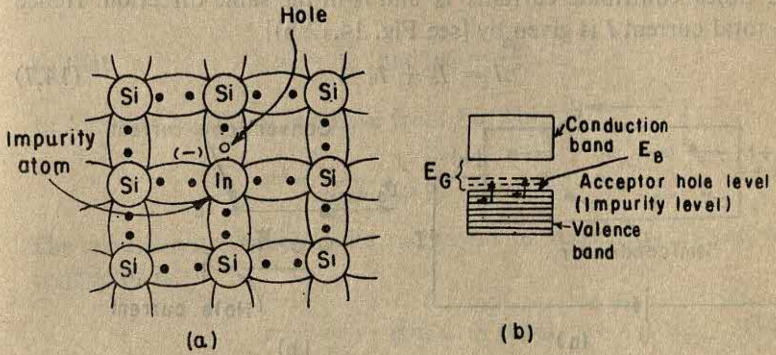


Fig. 14.11 *(a) Addition of indium (In) impurity to a silicon (Si) semiconductor (b) Excess holes introduce acceptor hole levels close to the valence band*

Notice from Fig. 14.11(b) that the hole energy level is at E_B above the top of the valence band. This means that if an energy E_B is supplied, the hole is free to move, i.e. there is one electron missing at the top of the otherwise full valence band. Since $E_B \ll E_G$, the conductivity of the doped *p*-type semiconductor is considerably increased. The reason is that, at room temperature, the holes in the valence band are free to move around because almost all the acceptor atoms are ionized due to thermal energy. Therefore the number density of holes (n_h) is equal to that of the acceptor ions (n_a). Further in *p*-type semiconductors n_h is greater than the electron number density (n_e), i.e.

$$n_a = n_h \gg n_e \tag{14.6}$$

Hence, in *p*-type semiconductors, holes are the *majority carriers* and electrons are the *minority carriers* of current.

14.7 ELECTRICAL RESISTIVITY OF SEMICONDUCTORS

We have seen that, in a semiconductor at room temperature, there are electrons in the conduction band and holes in the valence band. In *n*-type semiconductors the number density of electrons is much greater than that of the holes and in *p*-type semiconductors the opposite situation is obtained. We have also seen that, under an applied electric field, both holes and electrons move (in opposite directions); both contributing current in the direction of the field.

Suppose an electric field **E** is established in a block of a semiconductor (this can be done by connecting the ends of the block to a battery of voltage, say V) as shown in Fig. 14.12(a). Both electrons and holes contribute currents I_e and I_h in the same direction. Hence the total current I is given by [see Fig. 14.12(b)]

$$I = I_e + I_h \tag{14.7}$$

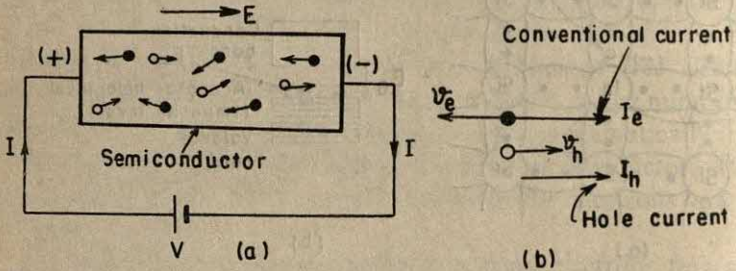


Fig. 14.12 *Drifting of electrons and holes in a semiconductor under an applied voltage V*

The electrons in the conduction band and the holes in the valence band drift randomly under the applied electric field just as electrons do in metals. Hence we can express the currents I_e and I_h in terms of the electron and hole drift velocities v_e and v_h just as we did in Chapter 3, Eq. (3.5). We can, therefore, write I as

$$I = n_e A e v_e + n_h A e v_h \tag{14.8}$$

where e is the magnitude of the electron charge and A is the cross-sectional of area of the block.

If the applied electric field E (hence the applied voltage V) is low, semiconductors obey Ohm's law

$$I = \frac{V}{R} \tag{14.9}$$

where R is the resistance of the semiconductor block. If l is the length of the block, then

$$|E| = \frac{V}{l} \tag{14.10}$$

Also the resistivity ρ of the block is given by

$$\rho = \frac{RA}{l} \tag{14.11}$$

Using Eqs (14.8) to (14.11) we have

$$\frac{E}{\rho} = e(n_e v_e + n_h v_h) \tag{14.12}$$

It is more useful to express the resistivity ρ in terms of a quantity μ, called *mobility* which is defined as drift velocity per unit electric field:

$$\mu = \frac{v}{E}$$

so that $$\mu_e = \frac{v_e}{E} \text{ and } \mu_h = \frac{v_h}{E}$$

In terms of μ_e and μ_h, we have from Eq. (14.12)

$$\rho = \frac{1}{e(n_e\mu_e + n_h \cdot \mu_h)} \tag{14.13}$$

The conductivity (which is the reciprocal of resistivity) of the semiconductor is

$$\sigma = \frac{1}{\rho} = e(n_e\mu_e + n_h\mu_h) \tag{14.14}$$

Thus we find that conductivity and resistivity of a semiconductor depend upon the electron and hole number densities and their mobilities. The mobilities of electrons and holes are found to decrease with increase in temperature. Also n_e and h_h increase with increasing temperature. The increase in n_e and n_h with temperature is so large that the conductivity σ increases with increase in temperature; the decrease in mobilities μ_e and μ_h with temperature being offset by the increase in n_e and n_h. Hence the resistivity of semiconductors decreases with increase in temperature. You will recall that the resistivity of metals (conductors) increases with increase in temperature.

EXAMPLE 14.1 The energy gap E_G of a pure silicon semiconductor is 1.1 eV. Its hole and electron mobilities respectively are

$$\mu_h = 0.048 \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1} \quad \text{and} \quad \mu_e = 0.135 \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1}$$

and are independent of temperature. Find the ratio between its conductivity at 400 K and 300 K. Assume that the temperature dependence of intrinsic carrier concentration is given by

$$n_i = n_0 \exp\left(-\frac{E_G}{2kT}\right)$$

where n_0 is a constant, T is the absolute temperature and $k = 1.38 \times 10^{-23}$ JK^{-1} is the Boltzmann constant.

Solution: For an intrinsic semiconductor $n_e = n_h = n_i$. Therefore, from Eq. (14.14) we have

$$\sigma = en_i(\mu_e + \mu_h) = en_i\mu$$

where $\mu = \mu_e + \mu_h = 0.135 + 0.048 = 0.183 \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1}$. Now

$$n_i = n_0 \exp\left(-\frac{E_G}{2kT}\right)$$

Therefore

$$\sigma = e\mu n_0 \exp\left(-\frac{E_G}{2kT}\right) = \sigma_0 \exp\left(-\frac{E_G}{2kT}\right) \qquad \text{(i)}$$

where $\sigma = e\mu n_0$, a constant independent of T. It is given that $E_G = 1.1$ eV and $k = 1.38\times10^{-23} \text{ JK}^{-1} = 8.12\times10^{-5} \text{ eVK}^{-1}$. From Eq. (i) we have

$$\frac{\sigma(\text{at } 400 \text{ K})}{\sigma(\text{at } 300 \text{ K})} = \frac{\exp\left(-\frac{E_G}{800\,k}\right)}{\exp\left(-\frac{E_G}{600\,k}\right)} = 280$$

The ratio is obtained by substituting the values of E_G and k and solving. Notice that the conductivity of an intrinsic semiconductor increases by a large factor with rise in temperature.

EXAMPLE 14.2 Germanium is doped 1 part in 10^6 with indium at room temperature. Find the conductivity of doped germanium. Given, concentration of germanium atoms $= 4.4\times10^{28} \text{ m}^{-3}$, intrinsic carrier concentration $(n_i) = 2.4\times10^{19} \text{ m}^{-3}$, $\mu_e = 0.39 \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1}$ and $\mu_h = 0.19 \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1}$.

Solution: Since the doping is 1 part in 10^6, the concentration of the acceptor atoms $= 4.4\times10^{28}\times1/10^6 = 4.4\times10^{22} \simeq n_h$, the concentration of holes. The hole concentration increases by an amount

$$\frac{n_h}{n_i} = \frac{4.4\times10^{22}}{2.4\times10^{19}} = 1.83\times10^3$$

Therefore, the new electron concentration is

$$n_e = \frac{n_i^2}{n_h} = \frac{2.4\times10^{19}\times2.4\times10^{19}}{4.4\times10^{22}} = 5.45\times10^{15} \text{ m}^{-3}$$

From Eq. (14.14) the conductivity is

$$\sigma = e(n_e\mu_e + n_h\mu_h)$$

Substituting the values of e, n_e, n_h, μ_e and μ_h, we get (on solving)

$$\begin{aligned}\sigma &= \sigma_e + \sigma_h \\ &= 3.41\times10^{-4} + 1.34\times10^{3} \\ &\simeq 1.34\times10^{3}\ (\Omega\text{m})^{-1}\end{aligned}$$

Notice that the contribution of electrons ($=3.41\times10^{-4}$) to conductivity is negligible as compared to that of holes ($=1.34\times10^{3}$). The conductivity of pure germanium is 2.18 $(\Omega\text{m})^{-1}$. Notice that the conductivity of germanium doped just 1 part in a million with indium has increased by a factor of about 10^3.

PART B SEMICONDUCTOR DEVICES

We have seen how the addition of impurities to a pure semiconductor gives rise to two types of charge carriers (electrons and holes) as well as a drastic increase in its conductivity. This fact has been used in the development of many semiconductor devices which use *p*- and *n*-type semiconductors. The first solid state transistor was made in 1947 by John Bardeen and William Brattain of the USA.

By themselves *p*- and *n*-type semiconductors find limited use. When the two types are combined to form a *pn* unit, a number of new characteristics appear which make the combination a very useful device, called the *pn junction diode*.

14.8 PN JUNCTION

When two semiconductors of *p* and *n* types are brought into contact, as shown in Fig. 14.13, they form what is called a *pn* junction.

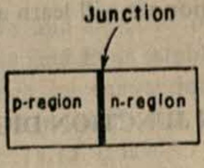


Fig. 14.13 *A pn junction*

A *pn* junction cannot be formed merely by placing a piece of *n*-type material against a piece of *p*-type material. A junction diode can be constructed in several ways. Two principal techniques of fabrication are: (i) *grown* junction and (ii) *fused* junction. In the grown junction technique, semiconductors are generally "grown" during their manufacture by dipping and withdrawing a "seed" from a crucible of molten semiconductor material (germanium or silicon) and then allowing the molten material (germanium or silicon) to cool and solidify on the seed. As the seed is continually withdrawn, more and more material solidifies and accumulates and so the crystal 'grows'. The molten semiconductor material is first given impurities of one kind, say for producing a *p*-type, and when enough is grown, impurities of opposite kind are added to change the type of material grown thereafter to the *n*-type.

In the fused junction technique, the *n*- and *p*-type silicon crystals are cut into thin slices called *wafers*. An aluminium film is placed on a wafer of *n*-type silicon and the combination is placed in an oven at a temperature of about 600°C. After several minutes, aluminium fuses to the surface of the semiconductor and produces *p*-type semiconductors immediately below the surface. Thus a *pn* junction is formed (Fig. 14.14).

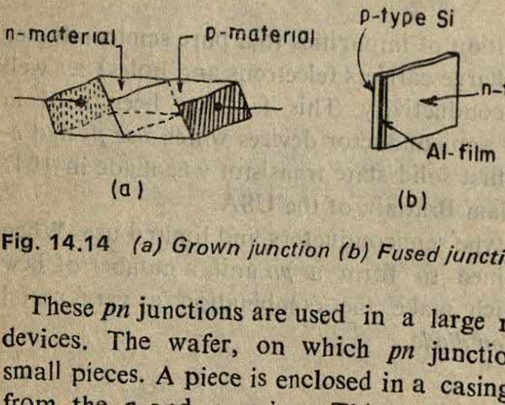


Fig. 14.14 *(a) Grown junction (b) Fused junction*

These *pn* junctions are used in a large number of semiconductor devices. The wafer, on which *pn* junctions are formed, is cut into small pieces. A piece is enclosed in a casing with leads coming out from the *p* and *n* regions. This is the simplest semiconductor device called the *pn* junction diode which is denoted by the symbol shown in Fig. 14.18(a). Another useful device is the *pnp* or *npn* transistor. In this section we will learn about the working and applications of these devices.

14.9 PN JUNCTION DIODE

In the *n*-region of a *pn* junction, the concentration of free electrons is higher than that of holes, whereas in the *p*-region, the concentration of holes is much higher than that of free electrons. Therefore, when a *pn* junction is formed, some electrons from the *n*-region (where their concentration is high) will diffuse into the *p*-region (where their concentration is low). Conversely, some holes from the *p*-region (where their concentration is high) will diffuse into the *n*-region (where their concentration is low). Sinc the hole is nothing but the vacancy of an electron, an electron diffusing from the *n* to the *p* region simply fills this vacancy, i.e. it completes the covalent bond. This process is called electron-hole recombination.

Depletion Region and Potential Barrier

As a result of electron-hole recombination, the electrons in the *n*-region (in the vicinity of the junction) are neutralized by holes, so that in this small region we are left with only ionized *donor* atoms (positive charges) as shown in Fig. 14.15. Similarly, the holes in the *p*-region (in the vicinity of the junction) are neutralized by electrons, leaving only ionized *acceptor* atoms (negative charges), as shown in Fig. 14.15. The positive and negative ions in a small region around the junction are bound and are, therefore, immobile. This small region in the vicinity of the junction which has been depleted of free charge carriers (electrons and holes) and has only immobile ions is called the *depletion region.*

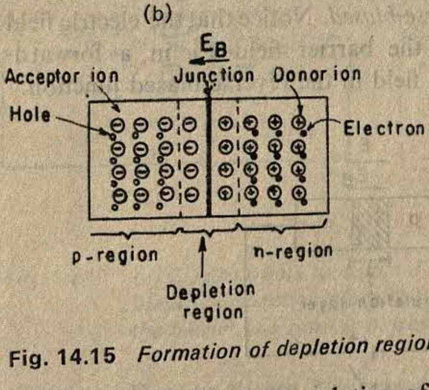


Fig. 14.15 *Formation of depletion region*

As a result of the accumulation of positive and negative immobile charges in the depletion region, an electric field $\mathbf{E}_{\mathrm{B}}$ is established in this region. The direction of the field (which by convention is the direction in which a positive charge will move in this field) is from the *n*-region to the *p*-region, as shown in Fig. 14.15. It is clear that the field $\mathbf{E}_{\mathrm{B}}$ will oppose any further flow of electrons from the *n*-region to the *p*-region and that of holes from the *p*-region to the *n*-region. Thus the electric field $\mathbf{E}_{\mathrm{B}}$ causes a *potential barrier* V_{B} at the junction which prevents further diffusion of electrons and holes into opposite sections of the depletion region. This happens because the overall negative charge of the *p*-region becomes sufficient to repel the free electrons in the n-region, keeping them from crossing the junction and vice versa. This limits the size of the depletion region.

Due to the existence of the potential barrier V_B across the junction, an electron from the n-region can cross the junction only if energy $\geqslant eV_B$ is supplied to it from outside. An equal amount of energy is required to move a hole from the p-region to cross over to the n-region across the barrier. The value of the barrier potential for germanium is about 0.3 V and for silicon about 0.7 V. The size of the depletion region is a few microns (1 micron $= 10^{-6}$ m). Assuming that $V_B = 0.5$ V and depletion width $= 10^{-6}$ m, the barrier field will be $E_B = 0.5 \text{ V}/10^{-6} \text{ m} = 5 \times 10^5 \text{ Vm}^{-1}$ which is a fairly high field at the junction.

Operation of a PN Junction: Forward and Reverse Bias

Let us now consider the *pn* junction connected to an external voltage source such as a battery. If the positive terminal of the battery is connected to the p-side and the negative terminal to the n-side, the junction diode is said to be *forward-biased* [Fig. 14.16(a)]. If the polarity of the battery is reversed, as shown in Fig. 14.16(b), the junction diode is said to be *reverse-biased.* Notice that the electric field **E** of the external source opposes the barrier field $\mathbf{E_B}$ in a forward-biased junction and helps the $\mathbf{E_B}$ field in the reverse-biased junction.

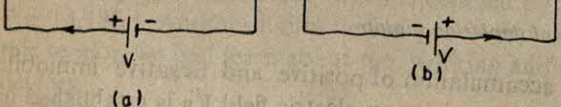


Fig. 14.16 *(a) Forward-biased pn junction, (b) Reverse-biased pn junction. Note that the depletion region becomes wider because* E *and* $\mathrm{E_B}$ *now add up*

Current in Forward-Biased PN Junction Diode

We have seen that the natural tendency of the majority carriers, free electrons in the n-section and holes in the p-section, is to combine at the junction. The point to note is that the combination of electrons and holes at the junction allows electrons to move in the same direction in both *p*- and *n*-sections.

With the *pn*-diode alone, the recombination of electrons and holes stops because of the potential barrier that builds up. If a battery is connected to the diode to overcome this potential barrier, the current flow can resume. If the polarity of the external battery is such that majority carriers in both the sections are driven towards the junction, then this type of biasing is called *forward bias*, causing forward the majority carriers to provide the current flow.

Forward bias is obtained with the negative terminal of the battery connected to the *n*-section and positive terminal connected to the *p*-side, because the negative terminal will repel free electrons in the *n* section towards the junction, and the positive terminal on the *p*-side will push the holes towards the junction. These free electrons and holes will then combine at the junction. But, for each combination, a free electron will enter the *n*-section from the battery, and a valence electron will leave the *p*-section to go to the battery. Therefore, the current will flow [Fig. 14.17(a)].

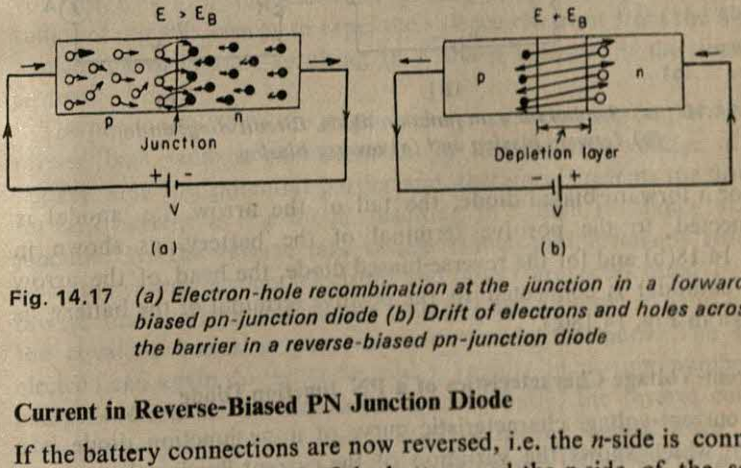


Fig. 14.17 *(a) Electron-hole recombination at the junction in a forward-biased pn-junction diode (b) Drift of electrons and holes across the barrier in a reverse-biased pn-junction diode*

Current in Reverse-Biased PN Junction Diode

If the battery connections are now reversed, i.e. the *n*-side is connected to the positive terminal of the battery and the *p*-side of the semiconductor is connected to the negative terminal, then electrons from the *n*-side will be attracted towards the positive potential and the negative potential will attract holes away from the junction, i.e. reverse bias repels the majority carriers from the junction. They cannot combine at the junction and majority current cannot flow when voltage with this polarity is applied. This process is called *reverse bias*. But reverse bias can cause a reverse current to flow, which is caused by the presence of minority carriers.

We know that although the *p*-section is doped to have excess holes, some electrons are freed because of thermal agitation. The free electrons in the *p*-section, and holes in the *n*-section, are called minority carriers. Now with reverse bias, these minority carriers are pushed towards the junction [Fig. 14.17(b)] where they recombine and allow electrons to enter and leave the *pn* diode in exactly the same way that the majority carriers do with forward bias. These minority carriers are much less in number than majority carriers; hence, the minority current is also much less than the forward majority carrier current.

Circuit Symbol for a PN Junction Diode

We use special symbols to represent electronic devices in an electrical circuit. The symbol for the simplest electronic device, namely the *pn* junction is shown in Fig. 14.18(a). The direction of the thick arrow is from the *p* to the *n*-region. The *p*-side is called the *anode* and the *n*-side is known as the *cathode*.

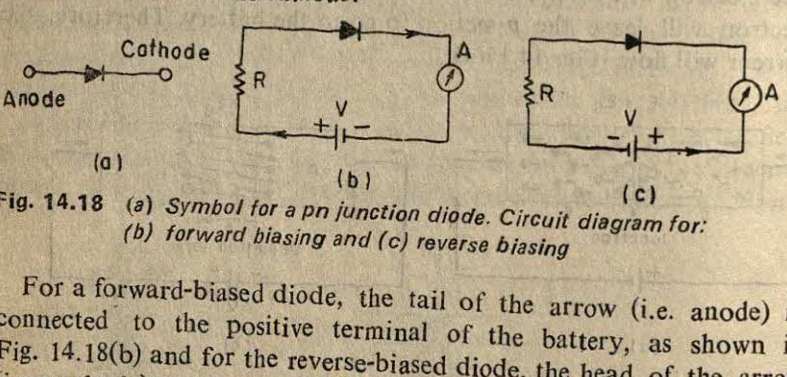


Fig. 14.18 *(a) Symbol for a pn junction diode. Circuit diagram for: (b) forward biasing and (c) reverse biasing*

For a forward-biased diode, the tail of the arrow (i.e. anode) is connected to the positive terminal of the battery, as shown in Fig. 14.18(b) and for the reverse-biased diode, the head of the arrow (i.e. cathode) is connected to the positive terminal of the battery, as shown in Fig. 14.18(c).

Current-Voltage Characteristics of a PN Junction Diode

The current-voltage characteristic curve of a *pn* junction diode is a graph which shows the variation of the current flowing through the junction with the change of voltage across the junction diode when it is forward-biased [Fig. 14.18(b)] and reverse-biased [Fig. 14.18(c)]. If the value of resistor $R = 0$, the voltage across the junction is V, the voltage of the battery. As R is gradually increased, the voltage across the junction is decreased. For each value of the voltage across the junction, the corresponding current is measured and the current-voltage graph is drawn. Figure 14.19 shows the current-voltage graph of a typical *pn* diode.

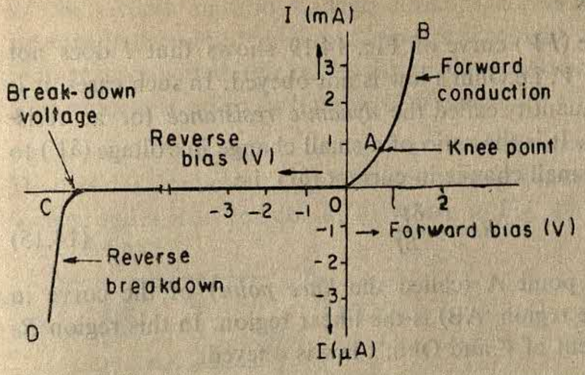


Fig. 14.19 *Current-voltage characteristic of a typical pn junction diode*

The portion OAB of the graph is the forward bias current versus the forward bias voltage characteristic. Notice how sharply the current increases with a small increase in voltage. Even with small applied voltages (~ 1V), the electrons passing across the junction gain sufficient kinetic energy to expel the valence electrons from the doped semiconductor atoms, resulting in a sharp increase in the forward current.

The portion OCD of the graph is the reverse bias current versus the reverse bias voltage characteristic. In this case the voltage of the battery aids the potential barrier and, therefore, prevents the flow of forward current in the diode, resulting in a high resistance at the junction. As the reverse bias voltage is increased, a stage is reached when the charge carriers (electrons and holes) are accelerated by the reverse bias voltage and acquire enough kinetic energy so as to break the covalent bonds in the semiconductor by collisions. The freed electron can again do the same and a rapidly increasing number of energetic electrons are produced. Consequently, the reverse current increases very rapidly and the material electrically 'breaks down'. The reverse biased voltage at which this happens is called the *breakdown voltage*. The value of the breakdown voltage varies from 1 or 2 V to several hundred volts depending on the number of density of the dopants and the width of the depletion region.

Note that in Fig. 14.19 the forward bias current is in milliamperes and the reverse bias current is in microamperes. The reason for choosing different current scales is that the former is much larger than the latter.

Dynamic Resistance

The current-voltage (*I-V*) curve of Fig. 14.19 shows that I does not vary linearly with V, i.e. Ohm's law is not obeyed. In such cases, it is useful to define a quantity called the *dynamic resistance* (or ac resistance) of the diode. It is the ratio of a small change in voltage (δV) to the corresponding small change in current (δI), i.e.

$$R_d = \frac{\delta V}{\delta I} \tag{14.15}$$

The region above point A (called the *knee point*) of the curve in Fig. 14.19 (i.e. the region AB) is the linear region. In this region R_d is almost independent of V and Ohm's law is obeyed.

PN Diode as a Rectifier

Since the *pn* diode conducts current more readily in one direction than in the other, it can be used to convert an alternating current into a unidirectional current. When an ac voltage is applied to a diode circuit, the diode will conduct relatively heavily when the polarity of the voltage produces a forward bias, but it will allow only a negligible current when the ac polarity reverses to produce reverse bias. As a result, current flows essentially for only one half cycle to produce a fluctuating dc at the output.

Half-wave rectifier circuit A half-wave rectifier consists of a transformer, a diode and a load resistor. The primary coil is connected to the ac mains, and the secondary coil to a load resistor R_L through the diode D as shown in Fig. 14.20(a).

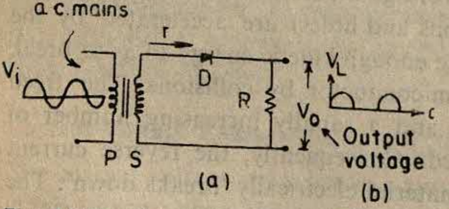


Fig. 14.20 *(a) Half-wave rectifier circuit*
(b) Pulsating output voltage

During the positive half-cycle of the input voltage sine wave, the diode is forward-biased and hence it conducts through R_L. The current flowing in the circuit produces a voltage across the load which has the same shape as the positive half-cycle of the input voltage V. During the next half-cycle (i.e. the negative half-cycle) of the sine wave, the diode is reverse biased. Hence, during this time, no current flows in the circuit and no voltage develops across R_L. Since only the positive half-cycle of the input appears across the load, the input ac voltage is converted into a pulsating dc voltage as shown in Fig. 14.20(b). This process is called *half-wave rectification.*

Full-wave rectifier circuit A full-wave rectifier consists of two diodes D_1 and D_2 connected to the centre-tapped secondary coil of a transformer and a load resistor R_L as shown in Fig. 14.21(a).

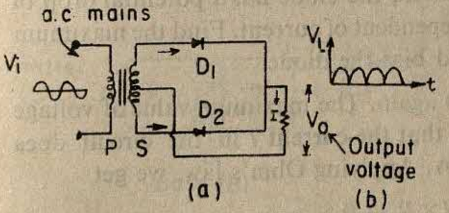


Fig. 14.21 *(a) Full wave rectifier circuit*
(b) DC output voltage

During the positive half-cycle of the input voltage sine wave, the diode D_1 is forward-biased and diode D_2 is reverse-biased, and the current flows in the direction shown in the figure. During the negative half-cycle, the diode D_2 is forward-biased and diode D_1 reverse-biased. Hence during this time also, a current flows through R_L in the same direction as in the previous half-cycle. Thus, the current flows (in the same direction) through R_L during the positive as well as the negative half-cycles of the input ac voltage. Hence the dc voltage shape across R_L is as shown in Fig. 14.21(b). This process is known as *full-wave rectification.*

Zener Diode as a Voltage Regulator

A diode especially designed to work only in the reverse break down region (region CD of the characteristic shown in Fig. 14.19) is called a Zener diode. The voltage across such a diode remains constant.

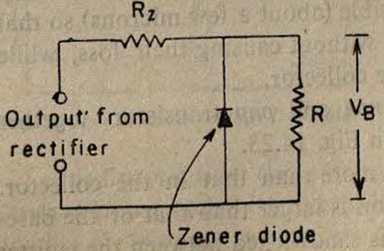


Fig. 14.22 *Circuit for a Zener diode as a voltage regulator*

Figure 14.22 shows a voltage regulator circuit using a Zener diode. Suppose the fluctuating (pulsating) voltage from the output of rectifier is applied across a Zener diode with a resistor R_Z as shown, this voltage will be divided between R_Z and the Zener diode. If this diode has a breakdown voltage $V_B = 5$ V, say, the voltage across the diode will always be $V_B = 5$ V, although the rectifier voltage and the voltage across R_z are both fluctuating. Thus, Zener diodes are used for making constant voltage power supplies, e.g. a battery eliminator.

EXAMPLE 14.3 A *pn* junction diode is designed to withstand currents up to a maximum of 10 mA. A resistor $R = 200\,\Omega$ is connected in series with it. When forward biased, the diode has a potential drop of 0.5 V which is assumed to be independent of current. Find the maximum voltage of the battery to forward bias the diode.

Solution: Refer to Fig. 14.18(b) again. The maximum value of voltage V of the battery should be such that the current I in the circuit does not exceed 10 mA $= 10 \times 10^{-3}$ A. Applying Ohm's law, we get

$$\begin{aligned} V &= I \times R + 0.5 \\ &= (10 \times 10^{-3}) \times 200 + 0.5 \\ &= 2.5 \text{ V} \end{aligned}$$

14.10 THE JUNCTION TRANSISTOR

The transistor is produced when another semiconductor element is added to the simple *pn* diode. The transistor is then composed of three semiconductor elements. The three elements are combined so that the two outer elements are doped with the same type of majority carriers, while the element that separates them has the opposite majority carrier. A transistor can be of two types: *npn* or *pnp* type.

The three elements of the transistor are the *emitter* (E) the *collector* (C) and the *base* (B). The emitter supplies the majority carriers for transistor current flow, the collector collects the current for circuit operation and the base controls the passage of electrons from the emitter to the collector. In fabricating the device, care is taken to make the base thickness as small as possible (about a few microns) so that it only controls the flow of electrons without causing their loss, while crossing the base region to reach the collector.

The schematic representation of *npn* and *pnp* transistors together with their circuit symbols is shown in Fig. 14.23.

The doping level in the emitter is more than that in the collector. The area of the base-collector junction is larger than that of the base-emitter junction. The base (which is sandwiched between the emitter and the collector) is thin and lightly doped. The reasons for these special features are explained below.

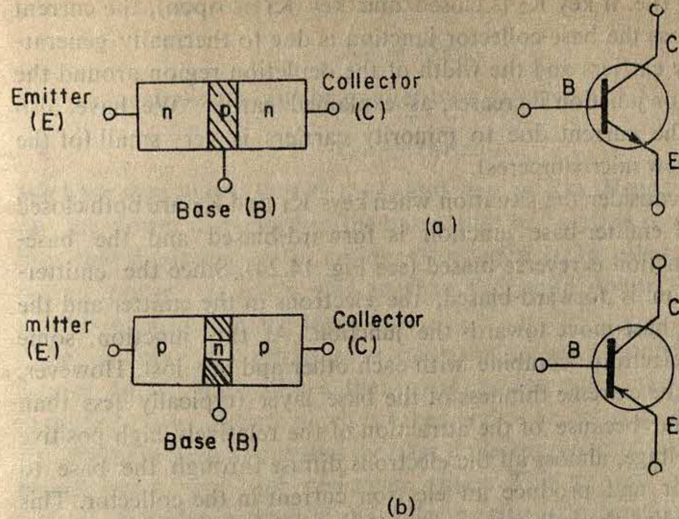


Fig. 14.23 *(a) npn transistor and its circuit symbol*
(b) pnp transistor and its circuit symbol

Transistor Action

Let us analyse the action of an *npn* transistor (Fig. 14.24). The transistor has two *n*-sections (emitter and collector) separated by a thin *p*-section base. The transistor can be regarded as two diodes joined back-to-back. The emitter-base junction of the left diode is shown forward-biased and the base-collector, of the right diode, is shown reverse-biased.

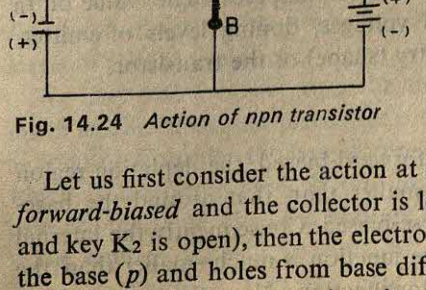


Fig. 14.24 *Action of npn transistor*

Let us first consider the action at the emitter-base junction. If it is *forward-biased* and the collector is left floating (i.e. if key K_1 is closed and key K_2 is open), then the electrons from the emitter (*n*) diffuse into the base (*p*) and holes from base diffuse into the emitter. On the other hand, if the base-collector junction is *reverse-biased* and the emitter is left floating (i.e. if key K_2 is closed and key K_1 is open), the current flowing across the base-collector junction is due to thermally generated minority carriers and the width of the depletion region around the base-collector junction increases, as explained earlier. We have also seen that the current due to minority carriers is very small (of the order of a few microamperes).

We now consider the situation when keys K_1 and K_2 are both closed so that the emitter-base junction is forward-biased and the base-collector junction is reverse-biased (see Fig. 14.24). Since the emitter-base junction is forward-biased, the electrons in the emitter and the holes in the base move towards the junction. At this junction, some holes and electrons combine with each other and are lost. However, because of the extreme thinness of the base layer (typically less than 10^{-3} cm) and because of the attraction of the relatively high positive collector voltage, almost all the electrons diffuse through the base to the collector and produce an electron current in the collector. This current is called the collector current (I_C) which is of the order of a few milliamperes. To make the collection of electrons efficient, the base-collector junction has a greater area than the emitter-base junction. Figure 14.24 shows the relative sizes of the two junctions. Notice that the area of the *pn* junction is larger than of the *np* junction.

Let us next consider the base current. This current is due to the small fraction of electrons which recombine in the thin base region. Since the base region is very thin, the base current (I_B) is a very small fraction of the collector current (about 1%). In other words about 99% of electrons pass through the base without recombining with holes. This means that the collector current (I_C) is only slightly less than the emitter current (I_E). In fact,

$$I_E = I_B + I_C$$

But $I_C \simeq I_E$, i.e. $I_B \ll I_E$ or I_C. In a typical transistor I_E and I_C are a few milliamperes and I_B is a few microamperes. The value of I_B depends on base thickness, bias voltages, doping levels of emitter, base and collector and the geometry (shape) of the transistor.

Amplifying Action of a Transistor

Since the emitter-base junction is forward-biased, the depletion region around this junction is much smaller than that around the base-collector junction which is reverse-biased. Consequently the internal resistance (R_{EB}) of the emitter-base junction is much smaller than the resistance (R_{BC}) of the base-collector junction. Now the power dissipation in the base-collector junction is

$$P_{BC} = I_c^2 R_{BC}$$

The power dissipated in the emitter-base junction is

$$P_{EB} = I_E^2 R_{EB}$$

We have seen above that $I_C \simeq I_E$ and $R_{BC} \gg R_{EB}$. Hence $P_{BC} \gg P_{EB}$, i.e. the power dissipated in the base-collector junction is much higher than that dissipated in the emitter-base junction. In other words, the output power is much greater than the input power. This is the amplifying action of the transistor.

For a detailed analysis of a transistor as an amplifier, refer to Section 14.13. We will see that input current, voltage and power are amplified at the output which is taken across an external load resistor R_L. We have seen above that the base-collector junction, being reverse-biased, has a very high resistance (R_{BC}). This permits a high load resistor R_L to be inserted in the collector circuit, without affecting the output current (I_C) and this is the basis of voltage amplification in a transistor. Thus $I_C^2 R_{BE} \ll I_C^2 R_L$ and a useful power gain is obtained. In fact the word transistor is a short form of the words *transfer resistor* indicating that the resistance of the source of the collector current is transferred from a small value (R_{EB}) to a large value (R_{BC}).

In a voltage amplifier (see Section 14.13), the input signal to be amplified is superposed on a steady voltage (V_{EB}) across the emitter-base junction. For a high voltage gain (which is defined as the ratio of output voltage to the input voltage), the change in the collector current (δI_C) should be large for a given change in the emitter-base voltage (δV_{EB}). Thus the quantity

$$g_m = \frac{\delta I_C}{\delta V_{EB}}$$

should be large for a high voltage gain. The quantity g_m, which has units of conductance, is called the *transfer conductance* (or simply *transconductance*) of the transistor; its value depends upon the doping levels of the three elements of the transistor and the biasing voltages.

We have stated earlier that the doping level of the base must be much smaller than that of emitter or collector. This can be understood as follows. Consider the hole current across the emitter-base junction which is forward-biased. If the doping levels in the emitter and the base are the same, then the hole (p) current flowing from the base (p) to the emitter (n) will be equal to the electron (n) current flowing from the emitter (n) to the base (p). Thus the electron-hole recombinations at their junction will be more, resulting in an increase in base current (I_B) and hence they reduce the collector current (I_C). The decrease in I_C reduces the power output ($I_C^2 R_L$) of the transistor, thus deteriorating the amplifying action of the transistor. Hence the base should be doped lightly.

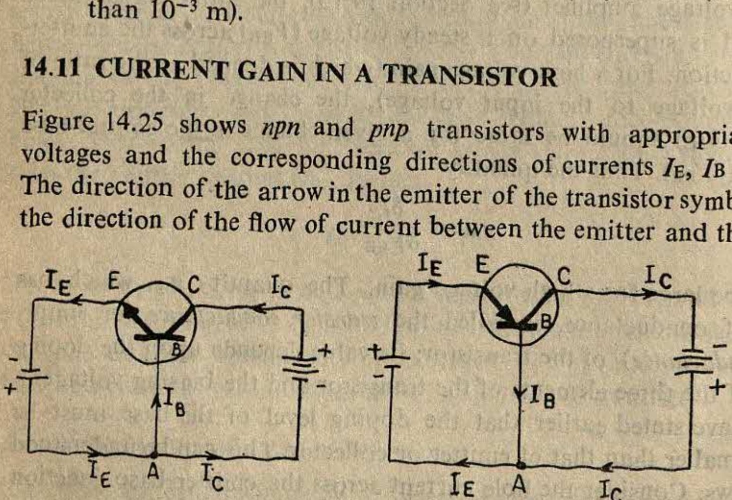
We have explained, in some detail, the action of an *npn* transistor. The action of a *pnp* transistor is similar to that of an *npn* transistor except that the majority current carriers (in the case of a *pnp* transistor) are holes instead of electrons and the bias votlages have opposite polarities so that the emitter-base junction is forward-biased and the base-collector junction is reverse-biased in both types of transistors.

The action of a transistor may be summarized as follows. For useful amplifying action, a transistor should have the following three features:

1. The three elements of the transistor, namely the emitter, the base and the collector should have different doping levels. The emitter should be heavily doped, the collector should be less heavily doped and the base should be lightly doped.
2. The three segments of the transistor are 'grown' in different sizes. The area of the base-collector junction is made larger than that of the emitter-base junction.
3. The thickness of the base layer should be very small (much less than 10^{-3} m).

14.11 CURRENT GAIN IN A TRANSISTOR

Figure 14.25 shows *npn* and *pnp* transistors with appropriate bias voltages and the corresponding directions of currents I_E, I_B and I_C. The direction of the arrow in the emitter of the transistor symbol gives the direction of the flow of current between the emitter and the base.



(a)

(b)

Fig. 14.25 *(a) npn and (b) pnp transistor circuits. The polarity of bias voltages is such that, in both cases, the emitter-base junction is forward-biased and the base-collector junction reverse-biased. Current directions are also shown*

Applying Kirchhoff's current law to junction A, we have

$$I_E = I_C + I_B \tag{14.16}$$

There are two current gains in a transistor, namely, α and β which are defined as

$$\alpha = \frac{I_C}{I_E} \tag{14.17}$$

$$\beta = \frac{I_C}{I_B} \tag{14.18}$$

The relation between α and β is obtained as follows. Equation (14.18) can be rewritten as

$$\begin{aligned}\beta &= \frac{I_C}{I_E}\cdot\frac{I_E}{I_B}\\ &= \frac{\alpha}{I_B/I_E}\\ &= \frac{\alpha}{(I_E - I_C)/I_E} \qquad (\because\ I_B = I_E - I_C)\\ &= \frac{\alpha}{1 - I_C/I_E}\end{aligned}$$

or

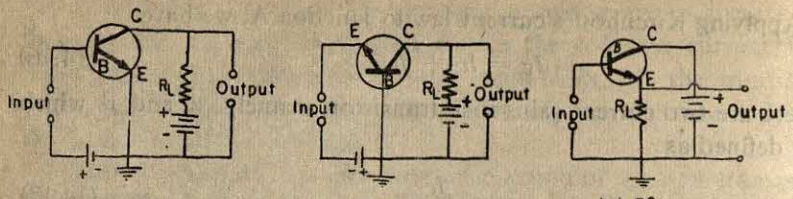
$$\beta = \frac{\alpha}{1-\alpha} \tag{14.19}$$

As stated earlier, I_C is only slightly less than I_E and I_B is much smaller than I_C (or I_E). Thus α is slightly less than unity (in a typical transistor $\alpha \sim 0.98$) and $\beta = \alpha/(1-\alpha)$ is large (50 or more).

Note: The definitions of α and β do not hold if both the junctions are reverse-biased or forward-biased. In the first case I_E, I_C and I_B are practically zero and in the second case, a large current flows as in a forward biased diode.

14.12 CURRENT-VOLTAGE CHARACTERISTICS OF A TRANSISTOR

A transistor is a three-element device. There can be three basic type circuits—common base, common emitter, and common collector [Figs. 14.26(a), (b) and (c)]. In all these three types of configurations, one element will always be common to the input and output circuits. Often the common lead is used as a reference point for the entire circuit, and is connected to the ground.



(a) CE (b) CB (c) CC

Fig. 14.26 *(a) Common-emitter (CE), (b) common-base (CB) and (c) common-collector (CC) connections for npn transistor*

We will restrict ourselves to the common-emitter transistor circuits [Fig. 14.26(a)] because they have the most wide ranging applications. However, the common-base and common-collector transistor circuits also have very useful applications.

Common-emitter Characteristics

The common-emitter characteristics of a transistor are its characteristics when the emitter is taken as the common reference terminal and is grounded (i.e. is at zero potential), the base is the input terminal and the collector the output terminal. These characteristics can be studied by using the common-emitter circuit shown in Fig. 14.27.

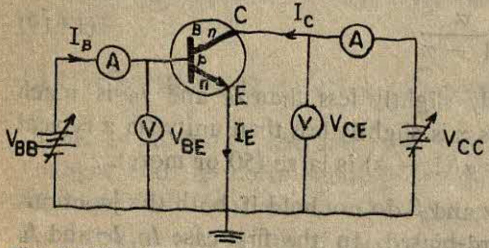


Fig. 14.27 *Circuit for studying the common-emitter characteristics of an npn transistor*

We study two types of characteristics—the input characteristics and the output characteristics. The graph of the base current I_B versus the base-emitter voltage V_{BE} keeping V_{CE} fixed is called the input characteristic [see Fig. 14.28(a)]. As long as V_{BE} is less than the barrier voltage, the current I_B is small, as in the case of a forward-biased diode. But when V_{BE} is increased to a value greater than barrier voltage, the current I_B increases sharply by small increases in V_{BE}. The value of I_B is, however, much smaller than that in a diode, because in the case of a transistor, more than 95% of the emitter electrons (in the *npn*

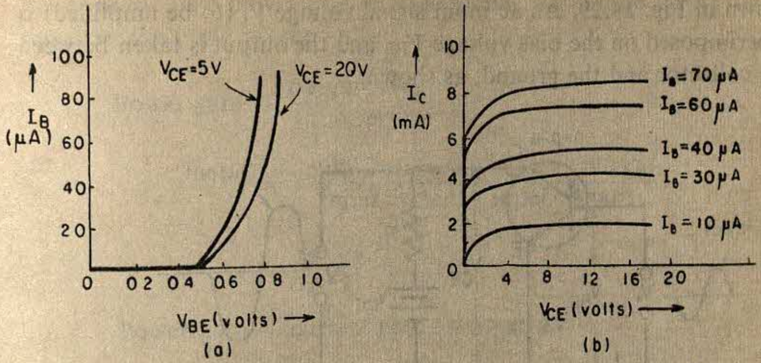


Fig. 14.28 (*a*) *Input characteristics, and* (*b*) *output characteristics of a common-emitter npn transistor*

transistor) or more than 95% of the emitter holes (in the *pnp* transistor) manage to reach the collector to constitute the collector current.

If the base-collector junction is reverse-biased, the input characteristics remain almost unaffected by small changes in V_{CE}. In such a situation, we can define the *input resistance* R_i as the slope at a given point on the characteristic curve, i.e.

$$R_i = \left(\frac{\delta V_{BE}}{\delta I_B}\right)_{V_{CE=\text{constant}}} \tag{14.20}$$

The output characteristic curves [shown in Fig. 14.28(b)], are obtained by plotting the collector current I_C as a function of V_{CE} for different values of the base current I_B. Notice that, for a given value of V_{CE}, I_C is larger for large values of I_B. Also notice that I_C is almost independent of V_{CE}, although it is controlled by I_B.

14.13 TRANSISTOR AS AN AMPLIFIER

The transistor amplifier is a circuit, consisting of at least one transistor, which is used for the amplification of voltage, current or power. To amplify means to increase the size of (or to magnify) an input signal. The output signal of an amplifier is an enlarged version of the input signal.

In order to use a transistor as an amplifier, the emitter-base junction is forward-biased and the base-collector junction reverse-biased. For the sake of explanation, we consider an *npn* transistor. In section 14.10, we discussed the action of this transistor. We describe now a common-emitter amplifier which is often used. The emitter lead is common to the input (emitter-base) circuit and output (base-collector) circuit, as shown in Fig. 14.29. An ac input signal voltage V_i (to be amplified) is superimposed on the bias voltage V_{BE} and the output is taken between the collector and the ground, as shown.

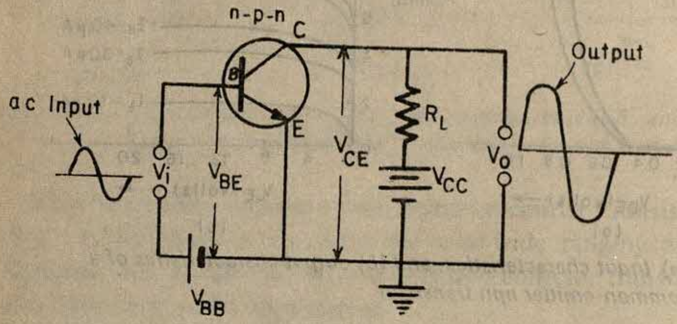


Fig. 14.29 *An npn transistor as an amplifier; common-emitter circuit*

Voltage Gain

The ratio of the output voltage V_0 to the input voltage V_i is called the voltage gain of an amplifier. In section 14.10 we have seen that the base-collector circuit (being reverse-biased) has a high resistance R_{BC} so that a high load resistor R_L can be inserted in the collector circuit, without affecting the output current I_C. This is the basis of voltage amplification by a transistor. Let $\alpha = I_C/I_E$ be the current gain of the transistor (see section 14.11). From Ohm's law we have

$$I_E = \frac{V_i}{R_{EB}}$$

where R_{EB} is the resistance of emitter-base circuit. Since the emitter-base junction is forward-biased, $R_{EB} \ll R_{BC}$ (see Section 14.10). The collector current is

$$I_C = \alpha I_E = \alpha \frac{V_i}{R_{EB}}$$

Therefore, the output voltage across R_L is given by

$$V_0 = I_C R_L = \alpha \frac{V_i}{R_{EB}} \cdot R_L$$

$$\therefore \quad \text{Voltage gain} = \frac{V_0}{V_i} = \alpha \frac{R_L}{R_{EB}} \qquad (14.21)$$

As explained earlier, R_L can be made much larger than R_{EB} and α is slightly less than unity. If $\alpha = 0.95$ and $R_L = 1000 R_{EB}$, the voltage gain will be 950. Thus, a high voltage gain is obtained in transistors.

Power Gain The power gain in an amplifier is the ratio of the output power to the input power, i.e.

$$\text{Power gain} = \frac{\text{output power}}{\text{input power}}$$

$$= \frac{\text{output current} \times \text{output voltage}}{\text{input current} \times \text{input voltage}}$$

$$= \frac{I_C}{I_E} \cdot \frac{V_0}{V_i}$$

$$\text{Power gain} = \text{current gain} \times \text{voltage gain}$$

EXAMPLE 14.4 If the input and output resistances in a common-base amplifier circuit are 400Ω and 400 kΩ respectively; (a) what is the voltage amplification when the emitter current is 2 mA and $\alpha = 0.98$? (b) Calculate the base and collector currents.

Solution:

(a) Emitter current = 2 mA

Input resistance = 400 Ω

Output resistance = 400 kΩ = 400×10^3 Ω

Input voltage = 2×400 mV = 800 mV

Collector current = $I_c = 0.98 \times 2 = 1.96$ mA

Output or collector voltage $= (1.96 \times 10^{-3}) \times (400 \times 10^{+3})$

$= 784$ V

$$\text{Voltage amplification} = \frac{784}{0.800} = 980$$

(b) Base current $I_B = I_E(1 - \alpha) = 2 \times 0.02 \text{ mA} = 0.04 \text{ mA}$

Collector current $= I_E - I_B = 2 - 0.04 = 1.96$ mA

EXAMPLE 14.5 The box in Fig. 14.30 represents an amplifier with an input internal resistance $R_i = 100$ Ω. It is connected to an ac voltage source via a resistor $R = 300$ Ω as shown. The voltage gain of the transistor is 40. If the peak-to-peak voltage of the input ac source is 5.0 V, find the peak-to-peak voltage of the output. What is the apparent voltage gain?

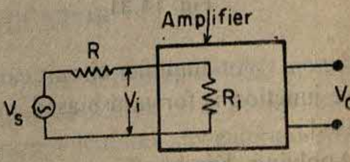


Fig. 14.30

Solution:

$$\text{Voltage gain} = \frac{V_0}{V_i}$$

$$V_0 = V_i \times \text{voltage gain}$$

If V_s is the voltage of the source, the voltage across the input of the amplifier is

$$V_i = \frac{V_s \times R_i}{R_i + R}$$

$$= \frac{5.0 \times 100}{(100 + 300)} = \frac{5}{4}\text{V}$$

$$\text{Output voltage } V_0 = V_i \times \text{voltage gain}$$

$$= \frac{5}{4} \times 400 = 500\ \text{V}$$

$$\text{Apparent voltage gain} = \frac{V_0}{V_s} = \frac{500}{5} = 100$$

EXAMPLE 14.6 In the circuit shown in Fig. 14.31, the base current $I_B = 5.0\ \mu\text{A}$, base resistor $R_B = 1.0 \times 10^6\ \Omega$, collector resistor $R_C = 1.1 \times 10^3\ \Omega$, the collector current $I_C = 5.0$ mA and the d.c. voltage in the collector circuit $V_{CC} = 6.0$ V. Can this circuit be used as an amplifier?

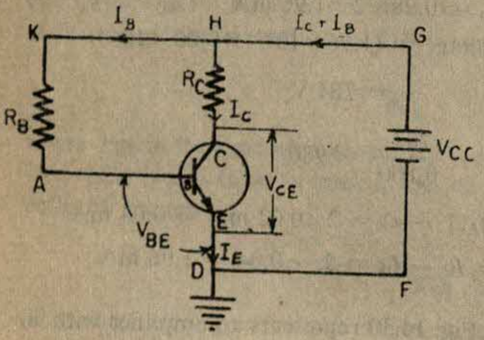


Fig. 14.31

Solution: An amplifier circuit can used as an amplifier if the emitter-base junction is forward-biased and the base-collector junction reverse-biased:

Applying Kirchhoff's loop rule to the loop ABEDFGHKA, we have

$$V_{CC} = V_{BE} + I_B R_B$$

where V_{BE} is the base-emitter voltage, which is

$$\begin{aligned} V_{BE} &= V_{CC} - I_B R_B \\ &= 6.0 - (5.0\times 10^{-6})\times(1.0\times 10^{6}) \\ &= 1.0 \text{ V} \end{aligned}$$

Applying Kirchhoff's law to the loop EDFGHCBE, we have

$$V_{CC} = V_{CE} + I_C R_C$$

where V_{CE} is the collector-emitter voltage which is

$$\begin{aligned} V_{CB} &= V_{CC} - I_C R_C \\ &= 6.0 - (5.0\times 10^{-3})\times 1.1\times 10^{3} \\ &= 0.5 \text{ V} \end{aligned}$$

The collector is +0.5 V with respect to the emitter and the base is +1.0 V with respect to emitter, so that the base is (1.0 − 0.5) = +0.5 V with respect to the collector. Therefore, both the emitter-base and base-collector junctions are forward-biased. Hence the circuit cannot be used as an amplifier.

EXAMPLE 14.8 If in Fig. 14.31 the value of $R_C = 400\ \Omega$, can the circuit then work as an amplifier. Take the rest of the data from Example 14.7.

Solution: The value of V_{CE} now will be

$$\begin{aligned} V_{CE} &= V_{CC} - I_C R_C \\ &= 6.0 - (5.0\times 10^{-3})\times 400 \\ &= 4.0 \text{ V} \end{aligned}$$

Now the collector is at +4.0 V with respect to the emitter and the base is at +1.0 V with respect to the emitter (see Example 14.7). Therefore, the base-collector junction is reverse-biased by (1.0 V - 4.0 V) = −3.0 V. Hence the circuit can now work as an amplifier.

14.14 TRANSISTOR AS AN OSCILLATOR

We have seen that in an amplifier, an externally applied ac input voltage is amplified at the output. An oscillator is a device which produces an output of a desired frequency without any external input voltage. This can be achieved if a part of the output of a transistor is fed back into the input circuit. If the feedback voltage is in the correct phase, oscillations of a constant amplitude can be obtained.

Figure 14.32 shows the basic constituents of a transistor oscillator. The oscillator circuit has two principal sections.

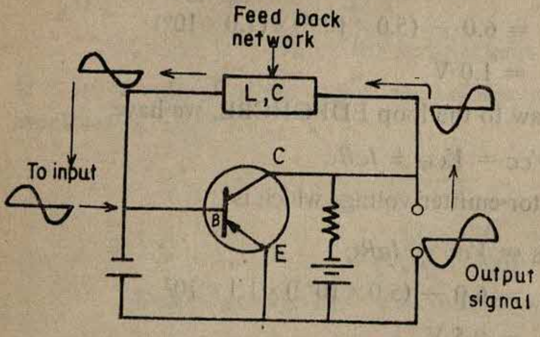


Fig. 14.32 *Common-emitter transistor oscillator (schematic)*

1. *Amplifying Section* This section is just a common-emitter amplifier having a high voltage gain (see section 14.13).

2. *Feedback section* It comprises a network of *LC* oscillatory circuits to provide the necessary positive feedback in the correct phase.

The input signal fed to the input circuit of the amplifier gets amplified at the output. Some portion of the output signal is fed to the feedback section which introduces a 180° phase shift in the signal. Exactly how this is achieved in the feedback section is beyond the level of the present discussion. The signal is then fed back to the input circuit. Thus the transistor supplies its own input and starts oscillating at a frequency which is determined by the values of the capacitances and inductances used in the feedback section. It is not possible, at this level, to describe the construction and the physics of the feedback section.

PART C: DIGITAL CIRCUITS

In Part B we have learnt some applications of semiconductor devices. We now discuss their applications in the most rapidly developing area of electronics—the area of computers. A computer is an information processing device. It can perform arithmetic operations (such as addition, subtraction, multiplication and division). It can also take logical decisions. It has a memory and can store information. The stored information may be used if and when desired. The computations are carried out at an extremely high speed and with fantastic accuracy and reliability. Modern computers can perform several million operations per second.

There are basically two types of computers—*analogue* and *digital*. The analogue computer uses *continuous* (unbroken) data and carries out computations with physical quantities, such as length, current, voltage, etc. Some examples of analogue devices are the slide rule (which uses length), the ammeter (which uses current) and the voltmeter (which uses voltage). Consider the ammeter. When a current is passed through it, it gives information about the current. The deflection of the needle gives a measure of the current passed. Current and deflection are both continuous quantities. The current is fed into the ammeter (input), it defects the current (processing of input) and gives a deflection (output). Thus, the ammeter is an analogue device. Analogue computers use this basic principle.

Digital computers, on the other hand, use *discrete* (discontinuous) data. Computations are carried out with discrete quantities, such as numerical digits. Some examples of digital devices are facit machines and electronic calculators. Digital computers are much faster and much more reliable and accurate than analogue computers. Modern computers are all digital computers.

Figure 14.33 schematically represents the basic organization of a computer system. It consists of a computer and supporting devices for input and output of data. The input device (card reader or teletypewriter) supplies the data to the computer. The data are the numeric and alphabetic forms which we use in everyday life. They are converted into the binary form (described below) with the help of electronic circuits called *encoders*. The computer processes the data and the results are obtained on the output device (printer) in the normal readable form with the help of electronic circuits called *decoders*.

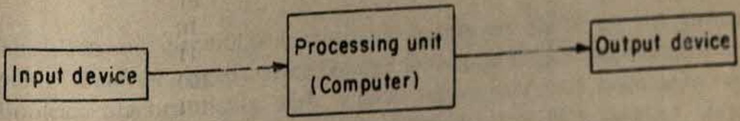


Fig. 14.33 *Computer system*

A computer is a digital system that performs various computational tasks. The word 'digital' means that the information in the computer is represented by variables that take a limited number of discrete values. These values are processed internally by the components of the computer that can maintain a limited number of discrete states. The decimal number system provides 10 discrete states corresponding to digits 0, 1, 2, 3, 4, 5, 6, 7, 8 and 9. We use these digits to represent a number in the decimal system. Each digit in a number has a place value. For example, consider a number 824 which is written as

$$824 = 8 \times 10^2 + 2 \times 10^1 + 4 \times 10^0 = 800 + 20 + 4$$

Digit 8 is in hundred's place, 2 in the ten's place and 4 in the unit's place. Here the base is 10. To represent the 10 digits of the decimal system, we need an 'object' which can take 10 different discrete states. Since such an object is difficult (though not impossible) to have, the decimal system is not used in computers. It is more convenient and reliable to use the binary number system for the simple reason that human logic tends to be binary (i.e. true or false, yes or no statements). The binary system is easy to implement, efficient and suited for high speed computations.

The binary number system uses two digits 0 and 1. A binary digit is called a *bit*. Binary numbers use the base 2 instead of the base 10 used in decimal numbers. For example, the binary number 1 stands for $1 \times 2^0 = 1$. The number 11 stands for $1 \times 2^1 + 1 \times 2^0 = 3$. Similarly, the binary number 1101 implies $1 \times 2^0 + 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 = 8 + 4 + 0 + 1 = 13$. Also the binary equivalent of a decimal number 45 is 101101. Thus, a string of 1's and 0's is used to represent a binary number. The group or string of bits could also represent a binary code for a letter of the alphabet and these letters can be used to construct different languages. Table 14.1 shows the binary equivalents of the first 15 decimal numbers.

Table 14.1 Binary equivalents of some decimal numbers

Decimal Number	Binary Number
0	00
1	01
2	10
3	11
4	100
5	101
6	110
7	111
8	1000
9	1001
10	1010
11	1011
12	1100
13	1101
14	1110

We will now learn the basic functions performed by semiconductor devices in practical circuits used in digital computers. Binary information is represented in a digital system by physical quantities called *signals.* Electrical signals, such as voltages (or currents), exist throughout a digital system in either one or two values and represent a binary variable equal to 1 or 0.

We have seen that, in amplifiers, the signal (voltage or current) is in the form of a continuous, time-varying voltage or current. Such signals are called *continuous* or *analogue* signals. Figure 14.34(a) shows a typical sinusoidal analogue signal. As stated earlier, digital computers do not use continuous signals; they use discrete binary signals. For example, a particular digital system may employ a signal of 3 V to represent a binary 1 or 0 V for binary 0. This is shown in Fig. 14.34(b).

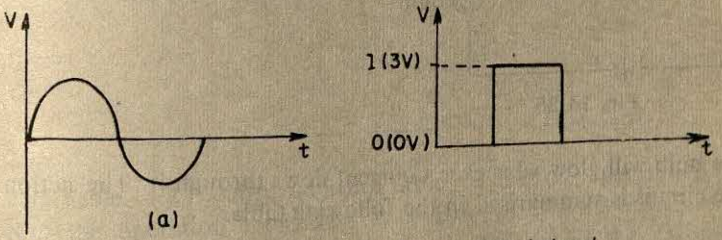


Fig. 14.34 *(a) An analogue signal (b) Binary digital signal*

In digital electronics, we use only two levels of voltage as shown in Fig. 14.34(b). Such signals are called *binary digital signals.* In digital circuits, only two values (represented by 0 and 1) of the input and output voltage are permissible.

14.15 BOOLEAN OPERATORS

The electronic circuits of a computer system are based on the principles of *Boolean* algebra after George Boole, an English mathematician. Boolean algebra deals with logical statements that have only two values, namely, either a *true* value or a *false* value. The logical statements are called *Boolean variables.* A Boolean variable must have a value that is either true or false. The true value of a Boolean variable is denoted by 1 and the false value by 0. It must be clearly understood that the symbols 1 and 0 have nothing to do with the numerics 1 and 0. In electronic circuits, the symbols 1 and 0 may be used to represent the active and passive states of a circuit component, say a switch, a diode or a transistor, as we will see now.

In ordinary algebra (using the decimal system) we use various mathematical operators, such as addition, subtraction and multiplication. In Boolean algebra, the three basic operators used are OR, AND and NOT.

The OR Operator

Consider the circuit shown in Fig. 14.35. S_1 and S_2 are two switches connected in parallel with a battery and a bulb.

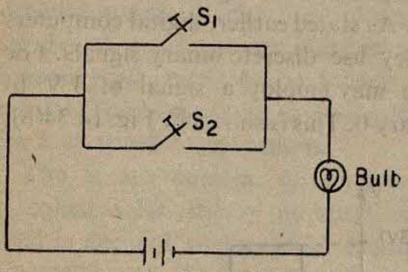


Fig. 14.35

The bulb will glow whenever a current flows through it. The action of the circuit is summarised in the following table.

Switch S_1	Switch S_2	Bulb glows
Open	Open	No
Open	Closed	Yes
Closed	Open	Yes
Closed	Closed	Yes

Now, let us agree to say that the open switch indicates false (F) and that the closed switch indicates true (T). Similarly, the glowing bulb represents true (T) and the non-glowing bulb represents false (F). Further let S_1 be called input A and S_2 input B. The state of the bulb may be called output X. The preceding table will now read as follows:

Input A	Input B	Output X
F	F	F
F	T	T
T	F	T
T	T	T

This table represents the OR operation. The OR operator connects two Boolean variables A and B to produce a new variable X. In Boolean algebra the OR operation is denoted by the symbol (+), i.e.

$$X = A + B$$

which is to be read as 'X equals A OR B'. Now since the true value of a Boolean variable is denoted by 1 and the false value by 0, the above table can be rewritten as follows:

Table 14.2 Truth table of OR operation

Input A	Input B	Output $X = A + B$
0	0	0
0	1	1
1	0	1
1	1	1

The Table 14.2 is called the truth table of the OR operation. Notice that the operation A OR B (= A + B) is used in the *inclusive sense*. The variable X (= A + B) is true (= 1) if either of the variables A or B or both A and B are true, i.e. if A = 1 and B = 0 or A = 0 and B = 1 or A = B = 1. The variable X (= A + B) is false (= 0) if both A and B are false, i.e. if A = B = 0.

The AND Operator

Now consider the circuit shown in Fig. 14.36. The switches S_1 and S_2 are now connected in series.

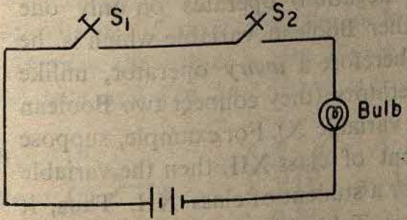


Fig. 14.36

The bulb will glow only if both the switches are closed. The action of switches is summarized in the following table.

Switch S_1	Switch S_2	Bulb glows
Open	Open	No
Open	Closed	No
Closed	Open	No
Closed	Closed	Yes

The operation of switches is called the AND operation. The AND operator also connects two input Boolean variables A and B to produce a new output variable X. The AND operation is denoted by the symbol (·), i.e.

$$X = A \cdot B$$

which is to be read a 'X equals A AND B'. Based on the above table, Table 14.3 will be the truth table of AND operation.

Table 14.3 Truth table of AND operation

Input A	Input B	Output $X = A \cdot B$
0	0	0
0	1	0
1	0	0
1	1	1

Notice that the variable X (=A.B) is true (=1) if and only if both the variables A and B are true, i.e. if $A = 1$ and $B = 1$. If one of them is false (= 0) then the variable X (= $A \cdot B$) is false (= 0).

The NOT Operator

The operator NOT (which means negation) operates on only one Boolean variable to produce another Boolean variable which is the negation of the first variable. It is therefore a *unary* operator, unlike OR and AND which are binary operators (they connect two Boolean variables A and B to produce a new variable X). For example, suppose the variable A = Ram is a student of class XII, then the variable NOT A (denoted by $\overline{A}$) = Ram is *not* a student of class XII. Thus, if $A = 1$, then $\overline{A} = 0$ and vice versa. Table 14.4 is the truth table of the NOT operation.

In electronic circuits, the above truth tables represent the states of various components in the circuit before and after the logical operations. The Boolean variables A and B are called *input variables* or

Table 14.4 Truth table of NOT operation

Input A	Output X = $\overline{A}$
0	1
1	0

simply input. The Boolean variable X representing $\overline{A}$, (A + B) and (A.B) is called the *output variable* or simply *output*.

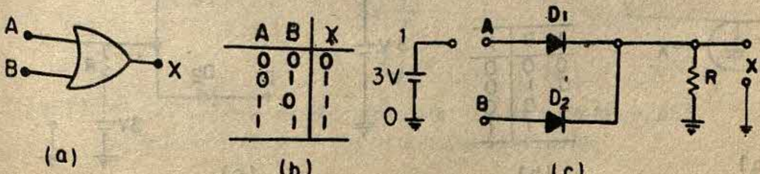
With this introduction to Boolean operations, let us now discuss how the three basic logic gates are realized by using semiconductor devices, such as diodes and transistors.

14.16 LOGIC GATES

Binary logic deals with binary variables and operations that assume a logical meaning. It is used to describe the binary information in a tabular or algebraic form. The manipulation of binary information is done by logic circuits called *gates. A gate is a digital circuit that follows a certain logical relationship between the input and output voltages.* They are, therefore, called *logic gates.* Various logic gates are commonly found in digital computer systems. There are three basic gates called OR, AND and NOT. Each gate is indicated by a graphic symbol and its operation (function) is described by means of an algebraic (*Boolean*) function. The input-output relationship of the binary variables for each gate is represented in tabular form in a *truth table.*

1. The OR Gate

Each gate has one or two binary input variables designated by A and B and one binary output variable designated by X. The OR gate has an output 1 if input A or input B or both inputs A and B are 1; otherwise the output is zero. Figure 14.37(a) is the logic symbol of the OR gate and Fig. 14.37(b) is its truth table. The algebraic symbol of



(a) (b) (c)

Fig. 14.37 *(a) Logic symbol of* OR *gate, (b) its truth table, and (c) realization of* OR *gate using diodes* the OR function is plus (+), i.e.

$$A + B = X$$

which is read as A OR B equals X. The OR gate may have more than two inputs and, by definition, the output is 1 if *any* input is 1.

An OR gate is realized as shown in the circuit of Fig. 14.37(c) using two *pn* diodes D_1 and D_2. The negative terminal of the battery is earthed (so that it is at zero potential) and corresponds to the 0 state and the positive (i.e. voltage 3 V in this example) is connected to 1 state. When both input terminals A and B are connected to 0, no current passes through either diode and, therefore no voltage develops across the resistor *R* and the output X is 0, i.e. $0 + 0 = 0$.

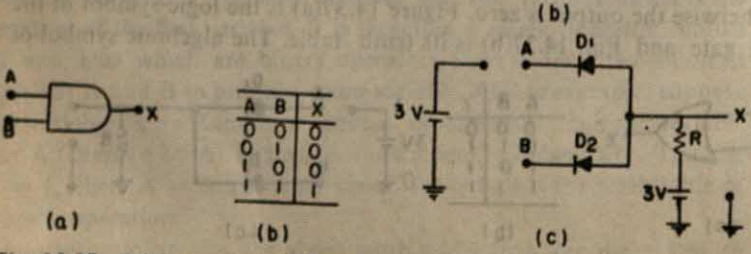
When input A is connected to 0 and B is connected to 1, the diode D_2 is forward-biased and hence conducts current. If the diodes are ideal, this current causes a voltage drop of 3 V across *R*, giving an output 3 V or X = 1, i.e. $0 + 1 = 1$.

When A is connected to 1 and B to 0, the diode D_1 conducts and the voltage drop across R is still 3 V, or X = 1, i.e. $1 + 0 = 1$. When the terminals A and B are connected to 1, both the diodes conduct, but the voltage across R cannot exceed 3 V; hence in this case also X = 1 i.e. $1 + 1 = 1$ (to be read as 1 OR 1 equals 1). Thus, the truth table is satisfied.

2. The AND Gate

It has two (or more) inputs and one output. The output X is 1 if input A AND input B are *both* binary; otherwise the output is 0. Figures 14.38 (a) and (b) show the symbol and the truth table of the AND gate respectively. The truth table shows that the output X is 1 only when both input A and input B are 1. The Boolean expression (i.e. the algebraic symbol) of the AND function is the same as the multiplication symbol of ordinary arithmetic. A dot is put between the variables thus

$$A.B = X$$



A	B	X
0	0	0
0	1	0
1	0	0
1	1	1

Fig. 14.38 *(a) Logic symbol of AND gate, (b) its truth table, and (c) realization of AND gate using diodes*

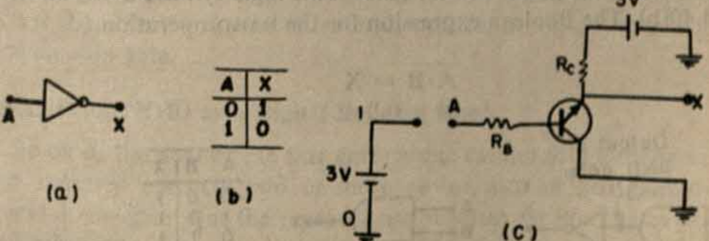
'Which is read as' A AND B equals X'. The dot in the AND operator is often dropped, so the statement A.B = X is often also written as AB = X.

As stated above, AND gates may have more than two inputs and, by definition, the output is 1 *if and only if all* inputs are 1.

The circuit in Fig. 14.38 (c) shows how two *pn* junction diodes can be used to realize the AND gate in practice. The resistor *R* is kept permanently connected to a 3 V battery. When both input terminals A and B are connected to 0, both diodes D_1 and D_2 conduct (because they are forward-biased) and voltage at output X will be zero. When A is connected to 0 and B to 1, the diode D_1 conducts but diode D_2 does not (because it is not forward-biased). The voltage output at X will be the voltage across D_1 which is zero, i.e. X = 0. The reason is that an ideal conducting diode has no voltage across it. When A is connected to 1 and B to 0, the diode D_2 conducts and the output is still zero (i.e. X = 0). When both A and B are connected to 1, neither D_1 nor D_2 conducts and the voltage output will be battery voltage, i.e. X = 1. Notice that the output is 1 when both the inputs are 1. If either input is 0, the output is 0. Thus the circuit obeys the truth Table 14.38 (b) of AND operation.

3. The NOT Gate

A NOT gate has only one input and one output. The output is 1 if the input is 0 and vice versa. The logic symbol and the truth table for a NOT gate are shown in Figs. 14.39(a) and (b) respectively. As stated earlier, the Boolean operation is represented as $\bar{A}$ = X which is read as 'NOT A equals X', i.e. X is the negation of A. In other words, X = 0 if A = 1 and X = 1 if A = 0.



A	X
0	1
1	0

Fig. 14.39 (*a*) *Logic symbol of* NOT *gate*, (*b*) *its truth table, and* (*c*) *realization of* NOT *gate using a transistor*

We cannot realize a NOT gate by using diodes. Instead a transistor has to be used. Figure 14.39(c) shows a simple circuit involving a transistor for realizing a NOT gate. The values of the base resistor R_B and the collector resistor R_C are chosen such that when a voltage of 3 V (which is the voltage of state 1) is applied at the base of the transistor, a large collector current flows, the voltage at the output X drops and the base-collector junction is forward-biased. When input A is connected to 0, the base-collector junction and the emitter-base junction are both reverse-biased. Hence the base current and the collector current are both zero. Consequently, the voltage at the output X is 3 V which is the voltage of state 1. When A is connected to 1, the transistor *saturates* and the voltage drop across R_C is almost equal to 3 V and the output X is 0 volt. Thus the truth table is satisfied.

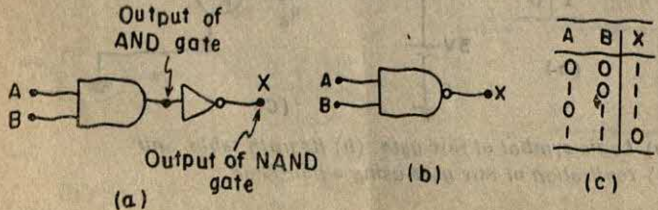
14.17 COMBINATION OF GATES

Complicated digital circuits use various combinations of the three basic gates OR, AND and NOT. Since the combinations are used repeatedly in such circuits, it becomes extremely cumbersome to show the same circuits of diodes, transistors and resistors. Instead, the universally accepted practice is to use only the symbols of the gates to represent more complicated gates. Two of the most commonly used combinations of gates are described now; they are called the NAND gate and the NOR gate.

The NAND Gate

The word NAND is an abbreviation of NOT-AND. A NAND gate is a combination of AND and NOT gates. If the output of an AND gate is connected to the input of a NOT gate, as shown in Fig. 14.40(a), the combined circuit is called a NAND gate whose logic symbol is shown in Fig. 14.40(b). The Boolean expression for the NAND operation is

$$\overline{A \cdot B} = X$$



A	B	X
0	0	1
1	0	1
0	1	1
1	1	0

Fig. 14.40 (*a*) *circuit of* NAND *gate* (*b*) *its logic symbol, and* (*c*) *its truth table*

which means that the output X here is just the negation of the output of the AND operation (whose truth table is shown in Table 14.3). Symbol $\overline{A \cdot B}$ is read as 'A AND B negated'.

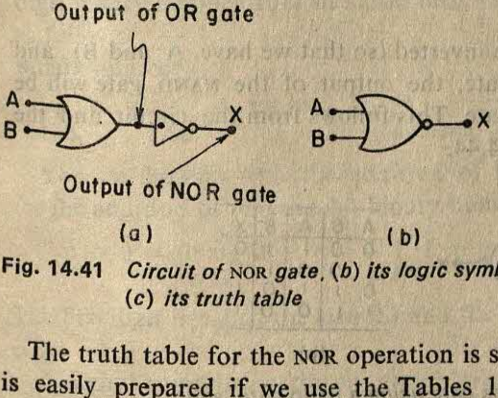
The truth table for the NAND gate can be easily prepared [see Fig. 14.40(c)] if we use the truth Table 14.3 for the AND operation and then find the negation of its output using the truth Table 14.4 for the NOT operation.

The NOR Gate

If we connect the output of an OR gate to the input of a NOT gate, as shown in Fig. 14.41(a), we get what is called a NOR gate (which is an abbreviation of NOT-OR). The logic symbol of the NOR gate is shown in Fig. 14.41(b). The Boolean expression for the NOR operation is

$$\overline{A + B} = X$$

which is read as 'A OR B negated'.



A	B	X
0	0	1
1	0	0
0	1	0
1	1	0

(a) (b) (c)

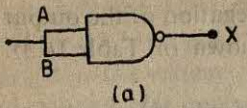
Fig. 14.41 *Circuit of* NOR *gate, (b) its logic symbol and (c) its truth table*

The truth table for the NOR operation is shown in Fig. 14.41(c). It is easily prepared if we use the Tables 14.2 and 14.3. The output (A + B) of the OR operation is simply negated to obtain the final output of the NOR gate.

NAND (or NOR) as a Digital Building Block

The OR or the AND or the NOT gates alone cannot give a different gate by repeated use, i.e. two or more OR or AND or NOT gates do not yield a new gate. But the repeated use of NAND (or NOR) gates can give all other gates like OR, AND and NOT. Hence the NAND (or the NOR) gate serves as a building block in digital circuits.

Figure 14.42 shows how we can get a NOT gate from a NAND gate. We simply join the two inputs A and B together to make one input. As is clear from the circuit, the truth table is as shown in Fig. 14.42(b). Since A = B, there is only one input and one output as in a NOT gate.

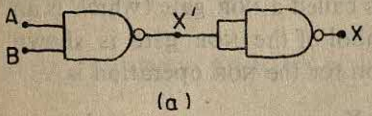


A	B=A	X
0	0	1
1	1	0

(b)

Fig. 14.42 *(a) Obtaining a* NOT *gate using a* NAND *gate. and (b) truth table*

Figure 14.43 shows how we can get an AND gate by using a NOT gate after a NAND gate. This is clear from the circuit and the truth table.

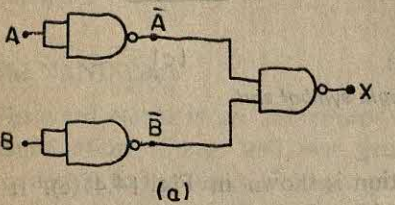


A	B	X'	X
0	0	1	0
1	0	1	0
0	1	1	0
1	1	0	1

(b)

Fig. 14.43 *(a) Obtaining an* AND *gate using a* NAND *gate, and (b) truth table*

If the inputs A and B are inverted (so that we have $\overline{A}$ and $\overline{B}$) and then applied to a NAND gate, the output of the NAND gate will be the same as that of an OR gate. This follows from the circuit and the truth table shown in Fig. 14.44.



A	B	$\overline{A}$	$\overline{B}$	X
0	0	1	1	0
1	0	0	1	1
0	1	1	0	1
1	1	0	0	1

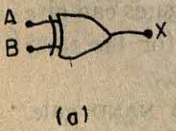
(b)

Fig. 14.44 *(a) Obtaining an* OR *gate using a* NAND *gate, and (b) truth table*

The XOR Gate

The XOR gate is an abbreviation of an *exclusive*-OR gate. The logic symbol and the truth table of an XOR gate are shown in Fig. 14.45. The circuit of the XOR gate uses combinations of two NOT gates and three NAND gates. It has two inputs and one output. The output X produced by the XOR gate is given by

$$X = A\overline{B} + \overline{A}B$$



A	B	X
0	0	0
0	1	1
1	0	1
1	1	0

(b)

Fig. 14.45 *(a) Logic symbol of* XOR *gate and (b) its truth table*

If we compare the truth table of the XOR gate with that of the OR gate (Table 14.2), we notice that the output in the XOR gate is 1 only when the inputs are different. Hence the name exclusive-OR. The XOR gates are used in binary adder circuits (see the following supplement).

SUPPLEMENT

BINARY ADDERS

A binary number is represented by two levels 0 and 1. Each digit of a binary number is called a *bit*. The first digit (i.e. the digit at the extreme right) is called the *least significant bit* (LSB) and the last digit (i.e. the digit at the extreme left) is called the *most significant bit* (MSB). You have already learnt how you can convert a binary number into its equivalent decimal number and vice versa. For example a four-digit binary number 1011 has a decimal equivalent which is obtained as follows.

Binary number

$$\underset{\text{MBB}}{1}\ 0\ 1\ \underset{\text{LSB}}{1} = \underset{\text{Conversion}}{1\times 2^3 + 0\times 2^2 + 1\times 2^1 + 1\times 2^0} = \underset{\text{Decimal number}}{11}$$

You are familiar with the addition of binary numbers. The rules for the addition of two one-bit binary numbers are as follows.

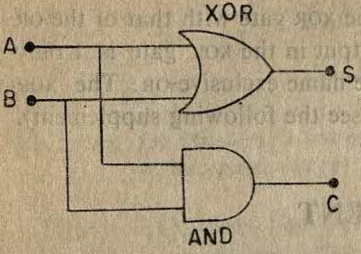
$$0 + 0 = \underset{(\text{c s})}{(0\ 0)},\ 0 + 1 = \underset{(\text{c s})}{0\ 1},\ 1 + 0 = \underset{(\text{c s})}{0\ 1},\ 1 + 1 = \underset{(\text{c s})}{1\ 0}$$

The first digit is called the *sum* (S) and the second is called the *carry* (C).

A digital computer performs only the operation of addition. The other operations such as subtraction, multiplication and division are performed by converting these operations into addition operations. The circuits performing the addition operations are called *adders*. The input variables A and B are called *addend* bits and the output variables are the *sum* (S) and *carry* (C).

The Half Adder

The circuit which performs the addition of two one-bit binary numbers is called the *half adder*. The two binary numbers A and B are the inputs variables. The sum(S) and the carry (C) are the output variables. It is essential to have two outputs because, as we have seen above, the addition of two one-bit binary numbers can yield two bits (e.g. 1 + 1 = 10). The logic circuit and the truth table of a half adder circuit are given in Fig. 14.46.



(a)

Inputs		Outputs	
A	B	C	S
0	0	0	0
0	1	0	1
1	0	0	1
1	1	1	0

(b)

Fig. 14.46 (*a*) *Circuit for a half adder, and* (*b*) *its truth table*

The truth table can be obtained by using the truth tables of XOR and AND gates. Notice that the sum S is 1, if and only if one of the inputs is 1; S is zero even when both inputs are 1. This is the case only with the XOR operation. Similarly the carry C is 1, if and only if both the inputs are 1 and it is 0 in all other cases. This is the case with the AND operation. Notice also that a two bit binary output can give the addition of up to three one bit binary numbers.

The Full Adder

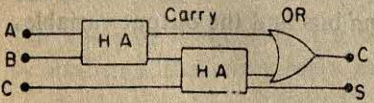
An adder circuit capable of adding three one bit binary numbers is called the *full adder*. The three bits to be added are the two inputs A and B and a carry C, which might have resulted by the addition of input bits. Here again two outputs C and S are essential because the bits to be added can be of the form

$$1 + 1 + 0 = 10$$

and

$$1 + 1 + 1 = 11$$

The symbol S stands for the sum and the symbol C is the carry bit to be carried over to the next step. The carry C is generated by the addition of A and B. Figure 14.47 shows the logic circuit and the



(a)

Inputs			Outputs	
A	B	C	C	S
0	0	0	0	0
0	0	1	0	1
0	1	0	0	1
0	1	1	1	0
1	0	0	0	1
1	0	1	1	0
1	1	0	1	0
1	1	1	1	1

(b)

Fig. 14.47 (*a*) *Circuit symbol for a full adder and* (*b*) *its truth table* truth table of a full adder circuit. It uses two half adders (H.A.) and an OR gate, as shown. The truth table is obtained from those of a half adder and the OR operation.

From the truth table, we notice that S is 1, if exactly one of the inputs is 1 or all the three inputs are 1. In all other cases, S = 0. On the other hand, C is 1 if either two inputs are 1 or all the three inputs are 1.

Integrated Circuits

We have seen that a half adder is made by using one AND and one XOR gate. An XOR gate is made by using two NOT, two AND and one OR gates. We have also seen that each of the NOT, AND and OR gates can be made by combining NAND gates. Thus, it turns out that a half adder will require eleven NAND gates. Each NAND gate is made using three resistors, two diodes and one transistor. Hence in a half adder, there will be 33 resistors, 22 diodes and 11 transistors. If they are connected in separate discrete circuits, the whole assembly will occupy a lot of space. With the phenomenal improvement in technology, it is now possible to make a complete circuit on a tiny silicon wafer. These circuits are called *integrated circuits* (IC's).

SUMMARY

The electrons in an atom are arranged in shells. The electrons in the outermost shell are called valence electrons. In an isolated atom, the electrons can occupy any of the allowed discrete energy levels. When a large number of atoms are brought together to form a solid, these energy levels spread into bands. The band structure of a material enables us to predict its ability to carry electric currents. In a conductor, either the valence band is partly filled or the valence and conduction bands overlap. In an insulator, there is a large forbidden band between the valence band and the conduction band. Materials having a small band gap between the valence and conduction bands behave as semiconductors.

Pure Si and Ge are the most well-known intrinsic semiconductors. Conduction is caused by electrons, freed from the covalent bonds by thermal agitation, and by the vacancies which are left behind. The controlled addition of impurities to a semiconducting material profoundly affects its electrical properties. Addition of pentavalent impurities to Ge on Si makes them *n*-type semiconductors. The majority carriers in these semiconductors are electrons. Semiconductors with holes as majority carriers are called *p*-type semiconductors. They are manufactured by the addition of trivalent impurities to Si or Ge.

Semiconductors have a wide range of practical applications as diodes, transistors, etc., and in a number of semiconductor devices. The simplest semiconductor device is a *pn* diode which combines a *p*-section containing an excess of holes with an *n*-section containing an excess of free electrons. The junction combination of electron holes causes a depletion region to be formed. Forward bias allows majority carrier current flow and reverse bias allows minority current flow. A *pn* diode converts alternating current into unidirectional current just as in a vacuum tube diode.

A transistor is produced when another semiconductor element is added to the simple *pn* diode. There are two types of transistor; *pnp* and *npn* transistor. Two outer *p*-sections are separated by an *n*-section semiconductor. The three elements of a transistor, the emitter, base and collector are comparable to the cathode, grid and plate of a vacuum tube. Transistors can be connected into three basic circuits—common-base, common-emitter, and common-collector. Alpha (α) is an expression of current amplification for a common-base circuit and beta (β) is called current amplification when the transistor is in the common-emitter mode. Transistors may be used as amplifiers and oscillators.

The fabrication of digital computers is the most rapidly developing area of electronics. The basic functions in these computers are performed by a combination of circuits involving semiconductor devices.

EXERCISES

A. Short-Answer Questions

1. Give one example each of amorphous and crystalline solids.
2. Name a conductor, an insulator and a semiconductor.
3. Distinguish between intrinsic and extrinsic semiconductors.
4. Name the current carriers in *n*- and *p*-type semiconductors.
5. What is a junction diode?
6. What is a transistor?
7. Distinguish between *pnp* and *npn* transistors.
8. Name the three basic logic gates.

B. Long-Answer Questions

1. Distinguish between crystalline and amorphous solids. Give examples of each type.

2. What do we understand by (a) a lattice, and (b) a unit cell? When is a unit cell said to be primitive? Is the choice of a unit cell unique?
3. Discuss how a two-dimensional lattice can be represented in mathematical terms.
4. What are the various possible types of plane lattice? What are their characteristics?
5. Discuss in detail the possible Bravais lattices in three dimensions. Give at least two examples of each type.
6. Discuss the various types of crystalline solids giving examples of each type. Discuss in detail the nature of the bonding in each case.
7. State, in each case, whether the following statements are true or false.
   - (a) Van der Waals forces represent the strongest type of bonding in solids.
   - (b) A close packed arrangement gives the maximum value of packing fraction.
   - (c) The choice of a unit cell is not unique.
   - (d) All alkali metals have a bcc structure.
   - (e) All noble metals have a simple cubic structure.
8. (a) What is a junction? How is it formed in a *pn* diode? What causes the potential barrier in a *pn* diode?
   - (b) When a *pn* diode rectifies an ac current, why does some current flow in both half-cycles?
   - (c) Draw the symbols for a *pnp* and *npn* transistor.
9. (a) Why is the emitter-base circuit forward-biased and the collector reverse-biased?
   - (b) Name the basic-type of circuits in which transistors can be connected. In which type of circuit configuration are very high power gains possible?
10. Draw and interpret the current–voltage characteristics of a *pn* junction diode. What is the break-down voltage?
11. How are junction diodes used as: (a) a half-wave rectifier, and (b) a full-wave rectifier?
12. Describe the action of a Zener diode as a voltage regulator.
13. Explain the action of a transistor. Why are the three elements made with different doping levels? Why is the base lightly doped? Why is the area of the base-collector junction made larger than that of the emitter-base junction? Why is the base layer made very thin?
14. Define the two current gains in a transistor and obtain the relation between them.
15. Draw and interpret the output characteristics of an *npn* transistor (common-emitter type).
16. Describe how a transistor is used as an amplifier. Obtain the expression for the voltage gain.
17. Name the three basic logic gates. Draw a circuit showing how each gate is realized. Prepare the truth table for each gate.
18. What are NAND and NOR gates. Draw their circuit symbols and obtain their truth tables.
19. How can a NAND gate perform the function of: (a) an AND gate, (b) a NOT gate, and (c) an OR gate.
20. What is the XOR gate? Where is it used ?

C. Multiple-Choice Questions

Choose the correct answer from the given alternatives.

1. When a semiconductor is subjected to an electric field, a current begins to flow through it which is composed of the movement of

   (a) only *n*-type carriers
   (b) only *p*-type carriers
   (c) both *n*- and *p*-type carriers
   (d) neither *n*- nor *p*-type carriers

2. Suitable impurities are added to a semiconductor depending on its use. This is done to

   (a) increase its life
   (b) enable it to withstand higher voltages
   (c) increase its electrical conductivity
   (d) increase its electrical resistivity.

3. The addition of a minute quantity of antimony to a silicon crystal makes it

   (a) a good insulator
   (b) a good conductor
   (c) a *p*-type semiconductor
   (d) an *n*-type semiconductor

4. To obtain an *n*-type semiconductor germanium crystal, it must be doped with foreign atoms whose valency is

   (a) 2 (b) 3 (c) 4 (d) 5

5. To obtain a *p*-type semiconductor germanium crystal, it must be doped with foreign atoms whose valency is

   (a) 2 (b) 3 (c) 4 (d) 5

6. The flow of current across a forward-biased *pn* junction is mainly due to

   (a) diffusion of charges
   (b) drift of charges
   (c) diffusion and drift of charges
   (d) diffusion of neutral atoms

7. A *pn* junction diode cannot be used

   (a) as a rectifier
   (b) for detecting light intensity
   (c) for getting light radiation
   (d) for amplifying an ac signal

8. To use a transistor as an amplifier

   (a) the emitter-base junction is forward-biased and the base-collector junction reverse-biased
   (b) both junctions are forward-biased
   (c) both junctions are reverse-biased
   (d) no biasing voltages are required.

9. An *npn* transistor is biased to work as an amplifier. Which of the following statements is not correct?

   (a) The electrons go from the base region to the collector region.
   (b) The electrons go from the emitter region to the base region.
   (c) The electrons go from the collector region to the base region.
   (d) The holes go from the base region to emitter region.

10. When the conductivity of a semiconductor is only due to the breaking of the covalent bonds, the semiconductor is called

    (a) donor (b) acceptor
    (c) intrinsic (d) extrinsic

11. A crystal lattice is

    (a) a random arrangement of atoms in a crystal
    (b) an ordered arrangement of atoms inside a crystal
    (c) a random arrangement of molecules but orderly arrangement of atoms
    (d) a piece of crystal

12. If the forward voltage in a diode is increased, the length of depletion layer will

    (a) decrease (b) increase (c) not change
    (d) increase proportional to applied voltage

13. If the electron concentration in a semiconductor is greater than the hole concentration, the semiconductor is

    (a) *n*-type (b) *p*-type
    (c) intrinsic (d) an insulator

14. To which logic gate does the truth table given below correspond?

	A	B	X
(a) NAND	0	0	0
(b) AND	0	1	0
(c) XOR	1	0	0
(d) OR	1	1	1

15. Digital circuits can be made by repetitive use of

    (a) AND gates (b) OR gates
    (c) NOT gates (d) NAND gates

D. Numerical Problems

1. If the input and output resistance in a common-base amplifier circuit are 400 Ω and 400 kΩ respectively: (a) What is the voltage amplification when the emitter current is 2 mA and $\alpha = 0.98$? (b) Calculate the base and collector current.
2. What is the voltage gain in a transistor if the emitter and collector currents are approximately equal? Assume that the emitting circuit impedance is 100 Ω and the collector impedance is 100 kΩ.

3. Given an *npn* transistor with a common-base connection and $\alpha - 0.98$, determine the base current and the collector current when the emitter current is 40 mA.
4. For a common-base amplifier, the input resistance is 800 Ω and output resistance is 600 kΩ.

   (a) Determine the voltage gain if the emitter current is 12 mA and $\alpha = 0.97$.
   (b) What is the power gain?
5. The power gain for a common-base amplifier is 800, and the voltage amplification factor is 840. Determine the collector current when the base current is 1.2 mA.

15

THE UNIVERSE

15.1 INTRODUCTION

From times immemorial, man has been watching with awe and amazement the motion of the various heavenly bodies of the universe. The branch of science that deals with the study of the universe is called *astronomy*. Unlike other sciences, astronomy is an observational rather than an experimental science. The reason is that the astronomers on earth cannot control the various phenomena taking place on the surface or in the interior of the stars and also in interstellar space. The conditions prevailing in the stars and in interstellar space are so diverse and so extreme that we cannot duplicate them in any laboratory on the surface of the earth.

The study of the universe began in ancient times when man observed the stars in the sky. On a clear night, several thousand stars are seen. If we observe the stars for a period of time, we find that they rise in the east, follow circular paths and set in the west, just as the sun does in the day. This led man to believe that the stars (including the sun) actually went round the earth. The Indian astronomer Aryabhatt (fifth century AD) was the first person to realize that it was the earth's rotation about an axis that caused the apparent circular motion of the stars.

Not all stars appear to move. The distant stars appear to remain fixed in space. This is due to the phenomenon called *parallax*. Suppose you are travelling in a train. If you observe two objects (such as trees) at different distances from the train, they seem to move relative to each other. The relative apparent motion in much less if the objects are very far away. It is for this reason that the distant stars do not seem to move. They are called "fixed stars".

Against the background of the fixed stars, a few heavenly bodies seem to move. They are the planets which in Greek means 'wanderers'. In order to explain the motion of planets, the great Greek astronomer Ptolemy (first century AD) proposed the *geocentric theory* according to which the earth was stationary at the centre of the universe

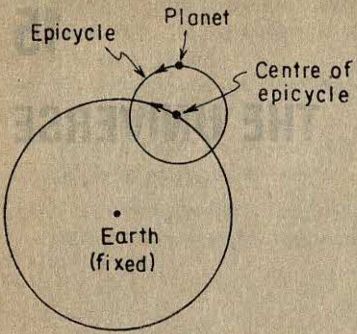


Fig. 15.1 *The geocentric model of planetary motion*

and the planets revolved around it. He assumed that each planet moved in a small circular path called the *epicycle*, the centre of which followed a circular path around the fixed earth (Fig. 15.1).

The geocentric theory reigned supreme for 1500 years until the great astronomer Copernicus (1473–1543) proposed his *heliocentric theory* according to which the sun was at rest and the planets (including the earth) revolved around it. This model of the solar system is accepted even today. Following this model, scientists like Galileo (1564–1642), Kepler (1571–1630), Newton (1642–1727) and many others made remarkable contributions towards the understanding of the universe.

It is indeed remarkable, as it is fascinating and reassuring, that some laws of nature which were discovered in laboratory experiments are valid everywhere in the universe. In other words, these laws are universal. Some examples of these laws are Newton's law of gravitation and the law of conservation of momentum and energy. These universal laws have been of great help in the understanding of the universe.

The electromagnetic radiations emitted by stars provide a link between the distant stars and earth. These radiations include γ-rays, x-rays, ultraviolet rays, the visible light, infrared radiation, microwaves and radiowaves, These radiations have to pass through the atmosphere of the earth to reach us. Most of these radiations, with the exception of those in the visible region and the near-infrared region, are absorbed by the atmosphere. But, with the advent of artificial satellites, it is now possible to receive these radiations and study them from above the earth's atmosphere. These studies give us very valuable information about many properties of the stars and planets such as their brightness, distances, masses, surface temperatures, composition, their motion and their evolution.

In this chapter, our aim is to discuss some methods used in the study of heavenly bodies, to present the available information concerning stars, galaxies, quasars, pulsars, etc. based on recent space probes and to discuss the possible theories of the origin of the universe.

15.2 IMPORTANT CONSTITUENTS OF THE UNIVERSE

There are three main constituents of the universe, which we consider one at a time.

The Solar system

The part of the universe in which we live is called the *solar system.* It includes the sun which is at the centre of the solar system along with planets (Mercury, Venus, Earth, Mars, Jupiter, Saturn, Uranus, Pluto and Neptune) which revolve around it in elliptical orbits. The solar system also includes the moon and the satellites of the planets and the asteroids, comets and meters.

Stars

Farther away from the solar systems are stars very much like the sun. The stars can occur as single stars, in pairs as binary stars, or in clusters like the Pleiades. On a clear night one can see about 2000 stars with the naked eye. But millions of stars can be seen with the help of telescopes. The distances between stars are enormous. Alpha Centauri which (after the sun) is the nearest star to us is at a distance of about 4 light years from the earth. One light year is the distance travelled by light in one year with a speed of 3×10^8 m s^{-1}. Thus, one light year is a distance of about 9.46×10^{15} m. The space between stars is almost completely empty, with traces of interstellar gas and dust.

Galaxies

A large group of stars is called a *galaxy*. We belong to a galaxy called the *Milky Way* which consists of a hundred billion stars including our Sun. There are billions of galaxies like the Milky Way. The galaxies nearest to earth can be seen with the naked eye. Very distant and faint galaxies can be seen with large telescopes. Photographs taken with these telescopes show millions of faint galaxies. The galaxies are the building blocks of the universe.

15.3 ASTRONOMICAL INSTRUMENTS

The instruments used in astronomical investigations can be classified into two groups: (i) instruments such as telescopes that collect visible light or radiowaves emitted by heavenly bodies, and (iii) instruments such as photographic plates, photocells, spectrograph, etc. which are the recording instruments. Telescopes that collect visible light are called *optical telescopes* while those that collect radiowaves are called *radio telescopes.*

A radio telescope consists of a large parabolic metal reflector which focusses the incoming radiowaves from space on an antenna placed at its focus. But unlike visible light radiowaves are not visible and they do not affect a photographic plate. The antenna is connected to a radio receiver where the signals are amplified. The output of the amplifier is recorded in a recorder. The output is in the form of a trace on a sheet of paper indicating the intensity of the incident radiation.

The radio telescope at Jodrell Bank in the UK has a reflector of diameter 75 m. The reflector of the radio telescope at Ootacamund in India is a cylinder of 500 m length and 50 m width. The largest radio telescope is in Puerto Rico and has a diameter of about 300 m.

Advantages of Radio Telescope Over Optical Telescope

1. Due to the enormous size of the reflector, a radio telescope can detect very faint radio signals. It can detect a very faint galaxy which no optical telescope can.
2. Radiowaves can pass through clouds and dust. Thus, a radio telescope can work even in cloudy conditions.
3. Radio telescopes are much cheaper than optical telescopes.
4. Radio telescopes can work as well during the day as they can during the night.

Recording Instruments

The most commonly used recording apparatus is the photographic plate. The permanent records of the image of the sun, the stars and the planets are used to measure the position, brightness and other features, such as the shape of the heavenly bodies. Photographic plates suffer from certain defects. The darkening of the plate is not proportional to the intensity of light. The comparison of intensities of light using photographic plates does not, therefore, give reliable information. Photoelectric cells are the most suitable for this purpose because the photoelectric current is proportional to the intensity of light. But with a photocell, one can observe only one star at a time. Hence, a combination of photographic and photoelectric techniques is used in practice. The recording of the amplified signals in a radio telescope is done on a piece of paper in a radio recorder.

15.4 THE STUDY OF THE SOLAR SYSTEM

In the early seventeenth century, Kepler formulated three laws which describe the motion of the planets around the sun in the solar system. The orbits of these planets are very nearly circular or more appropriately elliptical. These laws together with Newton's laws of gravitation are applicable to the motion of planets and other heavenly bodies. We shall now discuss how the physical properties, such as the distances of planets, their sizes, masses, etc., can be estimated on the basis of astronomical observations.

Distance

The following two methods can be used to estimate the distance of a heavenly body in the solar system.

1. *Triangulation Method* This method is the same as that used by a land surveyor and has been referred to in class XI. Suppose we wish to estimate the distance of, say, the moon M, from the earth (see Fig. 15.2). With the help of optical telescopes, the position of the moon is simultaneously observed from two widely separated places A and B on the surface of the earth. The position is ascertained with reference to stars which may be treated to be at rest. Thus, knowing the position of M from two different places, the angle AMB $(=\phi)$ can be easily found. This angle is called the parallactic angle.

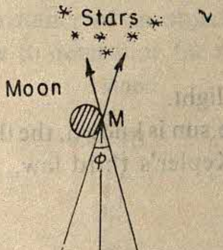


Fig. 15.2 *Triangulation method*

Now

$$\phi(\text{in radian}) = \frac{\text{Arc ACB}}{\text{Radius CM}} = \frac{\text{Arc(ACB)}}{\text{Radius(BM) of orbit}}$$

where $AM = BM = CM$

Thus $CM = \dfrac{ACB}{\phi}$

But $CM \simeq OM$ and arc $ACB \simeq$ chord AB. This is so because the distance of the heavenly body M is much larger than the radius of the earth. Thus

$$\mathrm{OM} = \frac{\mathrm{AB}}{\phi}$$

Knowing AB and ϕ the distance OM can be computed.

2. *Radar Echo Method* For nearer objects like the moon, the triangulation method gives fairly accurate results. For distant planets, this method does not give reliable results because the parallactic angle ϕ becomes too small to be measured accurately. In recent years, the distance of Venus has been accurately measured by the radar echo method. This method consists in sending radio signals to the planet. The waves get reflected from its surface. The reflected waves are picked up by sensitive receivers on earth. If t is the time taken for the waves to go to the planet Venus and come back, the distance S of the planet from the earth is given by

$$S = \frac{ct}{2}$$

where $c\,(=3\times10^8\ \mathrm{m\ s^{-1}})$ is the velocity of light.

Once the distance of any planet from the sun is known, the distances of all other planets can be obtained from Kepler's third law,

$$\frac{a^3}{T^2} = \text{Constant}$$

where a is the semi-major axis and T is the period of revolution. The distance of the earth from the sun has been calculated from the measured value of the distance of Venus from the sun as shown in Example 15.1.

EXAMPLE 15.1 The radar echo method gives the distance of Venus from the sun as 1.082×10^{11} m. The period of its revolution is 224.633 days. If the period of revolution of the earth around the sun is 365.257 days, estimate the distance of the earth from the sun.

Solution: From Kepler's third law we have

$$\frac{a^3}{T_E^2} = \frac{a_V^3}{T_V^2}$$

where

$$a_V = 1.082\times10^{11}\ \mathrm{m}$$

$$T_V = 224.633\ \text{days}$$

$$T_E = 365.257\ \text{days}$$

Hence

$$a_{\mathrm{E}}^3 = a_{\mathrm{V}}^3 \frac{T_{\mathrm{E}}^2}{T_{\mathrm{V}}^2}$$

or

$$a_{\mathrm{E}} = a_{\mathrm{V}}\left(\frac{T_{\mathrm{E}}}{T_{\mathrm{V}}}\right)^{2/3}$$

Substituting for a_{V}, T_{E} and T_{V}, we get

$$a_{\mathrm{E}} = 1.494 \times 10^{11} \text{ m}$$

Thus the distance of the earth from the sun is 1.494×10^{11} m. This distance is called the astronomical unit AU. Thus the distance of the earth from the sun is 1 AU where 1 AU $= 1.494 \times 10^{11}$ m. Similarly the distance of Venus from the sun is 0.723 AU.

Table 15.1 (column 2) at the end of this chapter gives distances of other heavenly objects.

Size

If the distance of an object like a planet or the moon is known, it is possible to determine the size of the object. Viewed through a telescope, the image of every heavenly object is a disc. The angle θ (in radian) between two extreme positions on the disc with respect to a certain point on the earth is determined with the help of a telescope (Fig. 15.3). The angle θ is called the angular diameter of the object.

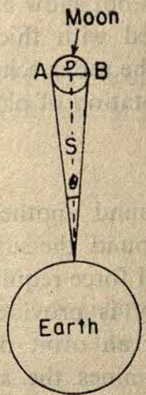


Fig. 15.3 *Size of the moon*

The linear diameter AB($=D$) of the object is then given by

$$\begin{aligned} D &= \text{distance} \times \text{angular diameter} \\ &= S\theta \end{aligned}$$

EXAMPLE 15.2 The angular diameter of the sun is 32 minutes of arc. If the distance of the sun from the earth is 1.5×10^{11} m, find the diameter of the sun.

Solution:

$$\begin{aligned} \text{Angular diameter } \theta &= 32 \text{ min} = \frac{32}{60} \text{ degree} \\ &= \frac{32}{60} \times \frac{3.14}{180} \\ &= 0.0093 \text{ rad} \end{aligned}$$

$$\text{Distance } S = 1.5\times 10^{11} \text{ m}$$

Linear diameter

$$\begin{aligned} D &= S\theta \\ &= 1.5\times 10^{11}\times 0.0093 \\ &= 1.4\times 10^{9} \text{ m} \end{aligned}$$

which is about 109 times the diameter of the earth. Table 15.1 (column 3) shows the sizes of some heavenly bodies.

Rotation

The earth revolves around the sun once in about 1 year and at the same time it rotates (spins) from west to east about its north-south axis once in about 24 hours. The other planets also rotate about their respective centres. The period of rotation of a planet is determined by observing the movement of the surface features (such as mountains, etc.) of the planets or that of cloud patches over the disc of the planet. Most planets (with the exception of Venus) rotate from west to east. The rotation of Venus is retrograde, i.e. from east to west. This method cannot work for Mercury which does not show any surface markings or for Venus which is always covered with thick clouds. Their period of rotation can be determined by the radar echo method.

Table 15.1 (column 4) shows the period of rotation of planets.

Mass

In the universe, one heavenly body revolves around another massive heavenly body. For example, planets revolve around the sun and the moon revolves around the earth. The centripetal force required by the lighter body to revolve around the heavier body is provided by the gravitational attraction between the two. For an orbit of a given diameter, the mass of the heavier body determines the speed with which the lighter body must revolve around it. Thus, if the period of revolution of the lighter body is known, the mass of the heavier body can be determined.

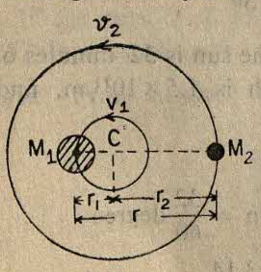


Fig. 15.4 *Motion about common centre of mass*

Suppose a lighter body (like the earth) of mass M_2 revolves round a heavier body (like the sun) of mass M_1 in a circular orbit of radius r (Fig. 15.4). Actually both bodies revolve around a common centre of mass C. If r_1 and r_2 are the radii of revolution of the two bodies about the centre of mass, then from the definition of centre of mass, we have

$$\frac{r_1}{r_2} = \frac{M_2}{M_1}$$

where $r_1 + r_2 = r$, as shown in the Fig. 15.4. Thus, we have

$$\frac{r_1}{r_2} + 1 = \frac{M_2}{M_1} + 1$$

or
$$\frac{r_1 + r_2}{r_2} = \frac{M_1 + M_2}{M_1}$$

or
$$(r_1 + r_2)M_1 = r_2(M_1 + M_2)$$

or
$$rM_1 = r_2(M_1 + M_2)$$

or
$$r_2 = \frac{rM_1}{M_1 + M_2} \tag{15.1}$$

Similarly
$$r_1 = \frac{rM_2}{M_1 + M_2} \tag{15.2}$$

If v_1 and v_2 are the speeds of the bodies M_1 and M_2 around their respective orbit and if T is the period of revolution, we have

$$v_1 = \frac{2\pi r_1}{T} \tag{15.3}$$

and
$$v_2 = \frac{2\pi r_2}{T} \tag{15.4}$$

Now let us consider the gravitational force acting on one of the bodies, say, M_2. The gravitational force F exerted on M_2 by M_1 is given by

$$F = G\,\frac{M_1 M_2}{r^2}$$

where G is the universal constant of gravitation.

This force provides the necessary centripetal force exerted on M_2 so as to revolve it in a circular path of radius r_2. The centripetal force

$$F = \frac{M_2 v_2^2}{r_2}$$

Thus, we have

$$G\,\frac{M_1 M_2}{r^2} = \frac{M_2 v_2^2}{r_2}$$

or
$$\frac{GM_1}{r^2} = \frac{v_2^2}{r_2} \tag{15.5}$$

Using Eq. (15.4) in Eq. (15.5), we have

$$G\frac{M_1}{r^2} = \left(\frac{2\pi r_2}{T}\right)^2 \cdot \frac{1}{r_2} = \frac{4\pi^2 r_2}{T^2}$$

Now substituting for r_2 from Eq. (15.1) in this equation, we get

$$(M_1 + M_2) = \frac{4\pi^2}{G} \cdot \frac{r^3}{T^2} \tag{15.6}$$

In the case of a planet moving round the sun, the mass M_2 of the planet is very small compared to the mass M_1 of the sun. In such cases, $M_1 + M_2 \simeq M_1$. Thus, from Eq. (15.6), we have

$$M_1 = \frac{4\pi^2}{G} \cdot \frac{r^3}{T^2} \tag{15.7}$$

Knowing the values of r, T and G; M_1 can be determined.

We notice that since the mass M_1 of the sun is constant, Eq. (15.7) gives for a planet

$$\frac{r^3}{T^2} = \text{Constant}$$

which is Kepler's law of periods.

EXAMPLE 15.3 The distance of the sun from the earth is 1.496×10^{11} m and the period of revolution of the earth around the sun is 365.257 days. Calculate the mass of the sun. Given $G = 6.668 \times 10^{-11}$ N m^2 kg^{-2}.

Solution:

$$r = 1.496 \times 10^{11} \text{ m}$$

$$T = 365.257 \times 24 \times 60 \times 60 \text{ s}$$

$$M_1 = \frac{4\pi^2}{G} \cdot \frac{r^3}{T^2}$$

Substituting the values of r, T and G, we get

$$M_1 = 1.99 \times 10^{30} \text{ kg}$$

Equation (15.7) can also be applied to a planet-satellite system in which M_1 will be the mass of the planet and M_2 that of its satellite.

Table 15.1 (column 5) gives masses of the planets. If M is the mass of a planet and r its mean radius, then the mean density of the planet and g, the acceleration due to gravity at its surface can be determined from the relations

$$\rho = \frac{3M}{4\pi r^3}$$

and

$$g = \frac{GM}{r^2}$$

The values of ρ and g for planets are given in columns 6 and 7 of Table 15.1.

EXAMPLE 15.4 The mass of the moon is 7.349×10^{22} kg and its mean radius is 1.738×10^6 m. (i) Calculate its mean density. (ii) What is the acceleration due to gravity at its surface? (iii) What will be the weight of a man on the moon, if he weighs 60 kg on the earth? Given $G = 6.668 \times 10^{-11}$ N m^2 kg^{-2}.

Solution:

$$M = 7.349 \times 10^{22} \text{ kg}$$

$$r = 1.738 \times 10^6 \text{ m}$$

(i) Mean density ρ of the moon is given by

$$\rho = \frac{3M}{4\pi r^3} = 3.34 \times 10^3 \text{ kg m}^{-3}$$

(ii) The acceleration due to gravity (g) on Moon's surface is

$$g_m = \frac{GM}{r^2} = 1.62 \text{ ms}^{-2}$$

(iii) Let g_e be the acceleration due to gravity on earth

$$g_e = 9.8 \text{ ms}^{-2}$$

$$\therefore \quad \text{Weight of man on Earth } (W_e) = mg_e$$

$$= 60 \times 9.8 = 588 \text{ N}$$

$$\text{Weight of the man on Moon } (W_m) = mg_m$$

$$= 60 \times 1.62 \text{ N}$$

$$= 97.2 \text{ N}$$

Thus, we find that the weight of a given body on earth is about six times the weight of the same body on the moon.

Surface Temperature

The radiations from the sun fall on the planets that move around it. The intensity of solar radiations on the surface of a planet depends on its distance from the sun, varying inversely as the square of the distance. Only a fraction on the incident radiation is reflected by the planet. The rest is absorbed by it. Planets shine by the reflected light of the sun. As a result of absorption of radiations, the surface of the planet gets heated. The planet then starts emitting radiation which lies mostly in the infra-red and radiowave region of the electromagnetic spectrum. By measuring the intensity of the radiation emitted by a planet, its surface temperature can be determined from Stefan's law of heat radiations, namely,

$$E = \sigma T^4$$

It is clear that the planets farther away from the sun will be colder than those closer to it because the intensity of radiation received by a planet varies inversely as the square of its distance from the sun. Mercury, being closest to the sun, is the hottest. Venus is an exception, because it has a very thick atmosphere of carbon dioxide which acts as a blanket and keeps it surface warm.

Atmosphere

The presence of an atmosphere around a planet or its satellite can be detected by observing the meteorological phenomena associated with clouds and storms. The composition of the atmosphere can be determined with the help of a spectrograph. Clouds are good reflectors of light. Hence they help to increase the reflecting power of a planet.. The reflecting power of a heavenly body is called *albedo.* The albedo of Moon is 0.06 which means that the surface of Moon reflects only 6% of the incident intensity, implying thereby that Moon has practically no atmosphere. Mercury also has a low albedo (~0.07) and, therefore, has no atmosphere. The albedos of Venus, Earth, Jupiter, Saturn, Uranus and Neptune are high which shows that they have thick atmospheres. Venus (owing to a thick blanket of clouds surrounding it) reflects about 80% of the incident radiation. Mars and Pluto have very thin atmospheres, their albedo being about 0.1.

As mentioned above, the nature of gases present in the atmosphere of a planet can be determined by studying the absorption spectrum of the radiation emitted by it with the help of a spectrograph. These studies have revealed that the atmosphere of Venus and Mars consist mostly of carbon dioxide and those of Jupiter and Saturn contain hydrogen, helium, methane, etc. The atmosphere of the earth consists of 80% nitrogen, 20% oxygen with some carbon dioxide and water vapour.

Other Objects in the Solar System

Besides the planets and their satellites, the solar system includes objects called asteroids, comets, meteors and meteorites.

Asteroids Asteroids are a group of small objects revolving round the sun between the orbits of Mars and Jupiter. More than 2000 asteroids have been discovered. The largest among them has a diameter of about 700 km and is called Ceres. It circles the sun once every 4½ years. The smallest asteroid may have a diameter of only about 100 m. They are pieces of rock which are believed to have broken away from planets, possibly due to the gravitational effect of Jupiter.

Comets A comet consists of a small mass of rock-like material surrounded by large masses of substances, such as water, ammonia and methane. These substances are easily vaporized. Comets have highly elliptical orbits around the sun. When a comet is very far away from the sun, it shows no tail. As it heads towards the sun, it begins to get elongated in the direction away from the sun as shown in Fig. 15.5. It then shows a tail which always points away from the sun, regardless of its direction of motion.

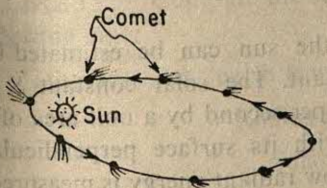


Fig. 15.5 *The orbit of a comet*

You will learn in higher classes that radiation exerts pressure like a gas does. This is called the radiation pressure. The formation of the tail of a comet is due to radiation pressure. As the comet approaches the sun, some of its substances are vaporized by the heat of the sun. The pressure exerted by the light of the sun on these vapours forces them away from the comet.

Meteors and Meteorites If you scan the sky on a clear moonless night, you might see what is called a 'shooting star' or a meteor. Actually they are not stars but small pieces of rock or metal. Occasionally, the earth in its orbital motion around the sun, comes so close to one of these objects that the object is attracted towards the earth by the large gravitational pull exerted by the earth. As it enters the earth's atmosphere it is accelerated to extremely high speeds.

Due to the friction of the air, the object is heated to very high temperatures and it begins to glow and gets completely burnt before it reaches the earth. These objects blazing through the atmosphere are called *meteors.* The larger objects may survive the heat produced by friction and may not be completely burnt. These blazing objects which manage to reach the earth are then called *meteorites.*

15.5 THE SUN

The sun is our nearest star. We shall see in Sec. 15.6 that the sun is a star of average size, mass and brightness. It looks big and bright only because it is much closer to us than the other much bigger and brighter stars. We have already seen how the distance and the mass of the sun are determined. The distance of the sun from the earth is 1.496×10^{11} m which is called the astronomical unit (AU). The mass of the sun is 1.989×10^{30} kg. The diameter of the sun is about 1.39×10^9 m which is approximately 109 times that of the earth. We shall now consider other important properties such as luminosity, temperature, etc.

Surface Temperature of the Sun

The temperature of the surface of the sun can be estimated by measuring what is called the solar constant. The solar constant S is the amount of radiant energy received per second by a unit area of a perfectly black body placed on earth with its surface perpendicular to sun's rays. In class XI we learnt how radiant energy is measured. Experimental measurements have shown that the value of the solar constant is 1.39×10^3 W m^{-2}. This means that a surface area of 1 m^2 of a black body on earth held perpendicular to the sun's rays receives 1.39×10^3 J of radiant energy per second.

The total radiant energy emitted by the sun is, however, much larger. The earth receives only a small fraction of it which is directed on it. To calculate the total radiant energy, let us imagine a sphere of radius R equal to 1 astronomical unit (i.e. the distance between the sun and the earth) with the sun at its centre (Fig. 15.6). The sun radiates energy in all directions. Now, if S is the solar constant, the total amount of radiant energy E received per second by the sphere of radius R is given by

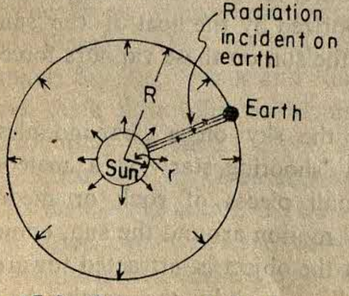


Fig. 15.6 *Calculation of the surface temperature of the sun*

$$\begin{aligned} E &= \text{surface area} \times \text{solar constant} \\ &= 4\pi R^2 S \end{aligned} \tag{15.8}$$

Assuming that there is no loss of energy as it travels, the energy emitted by the sun per second must also be E. If T is its surface temperature, then, according to Stefan's law, the energy emitted by a unit area of the sun per second, assuming the sun to be a black body is given by σT^4 where σ is Stefan's constant. Thus, if r is the radius of the sun, the total radiant energy E emitted by the sun per second is given by

$$E = 4\pi r^2 \sigma \cdot T^4 \tag{15.9}$$

Equating Eqs (15.8) and (15.9) we have

$$T = \left(\frac{R}{r}\right)^{1/2} \left(\frac{S}{\sigma}\right)^{1/4} \tag{15.10}$$

Knowing R, r, S and σ the absolute temperature T of the surface of the sun can be determined.

Example 15.5 Compute the surface temperature of the sun from the following data:

Distance of the earth from the sun (R) = 1.496×10^{11} m

Radius of the sun (r) = 6.928×10^8 m

Solar constant (S) = 1.39×10^3 W m^{-2}

Stefan's constant (σ) = 5.73×10^{-8} W m^{-2} K^{-4}

Solution: Using Eq. (15.10) above and substituting the values of R, r, S and σ, we have

$$\begin{aligned} T &= \left(\frac{R}{r}\right)^{1/2} \left(\frac{S}{\sigma}\right)^{1/4} \\ &= 5800 \text{ K} \end{aligned}$$

Example 15.6 Calculate the energy radiated per second by the sun. Use the data given in Example 15.5 above.

Solution:

$$\begin{aligned} E &= (4\pi r^2)S \\ &= 4\pi \times (1.496 \times 10^{11})^2 \times 1.39 \times 10^3 \\ &= 3.9 \times 10^{26} \text{ W} \end{aligned}$$

This is called the solar luminosity.

Condition at the Centre of the Sun

The temperature calculated above is the temperature of the outermost layer of the sun. This layer is called *solar corona* or *chromosphere* which is transparent to visible light. The disc which we actually see when we look at the sun lies below the chromosphere and is called the *photosphere* which produces light. As we go deep below the surface layers of the sun, the temperature, pressure and density go on increasing due to the compression caused by the weight of the layers above. Calculations show that the temperature at the centre of the sun is as high as 14×10^6 K. At such high temperatures, no substance can exist in a solid or liquid form. Hence the sun is gaseous all the way. About 70% of the mass of the sun is hydrogen, about 28% helium and the rest only 2% is composed of other heavier elements from lithium to uranium. The density of the sun increases from about 10^{-4} kg m^{-3} at the surface to about 10^4 kg m^{-3} at the centre. The pressure at the centre of the sun is of the order of 2×10^{16} Nm^{-2}.

Energy Production in the Sun (Source of Solar Energy)

What is the source of energy of the sun? This question baffled scientists for a very long time. There must be a source within the sun steadily supplying the energy it radiates. In the 19th century, Helmholtz and Kelvin suggested that the gravitational energy liberated during the slow contraction of the sun was the source of energy. Let us compute this energy. We assume, for simplicity, that the sun is a sphere of uniform density. It can be shown by a detailed calculation that a uniform sphere of mass M and radius R possesses gravitational potential energy $= -3GM^2/5R$, where G is the gravitation constant. This energy is due to the work done in bringing each particle of the sphere from infinity to form the sphere. If the sphere shrinks from infinite radius to radius R, the gravitational energy released in this process is

$$E_g = \frac{3GM^2}{5R}$$

Substituting the values of G, M and R, we obtain

$$E_g = 2.4 \times 10^{41} \text{ J}$$

In Ex. 15.6 we have seen that the energy radiated per second by the sun is

$$E = 3.9 \times 10^{26} \text{ J s}^{-1}$$

Assuming that the sun has been radiating energy at this rate since its birth, its age would be

$$\frac{E}{E_g} = \frac{2.4 \times 10^{41}\ \mathrm{J}}{3.9 \times 10^{26}\ \mathrm{J\ s^{-1}}} \simeq 6 \times 10^{14}\ \mathrm{s}$$

which is about 20 million years. This is much smaller than the age of the earth (about 4 billion years or more) as found by geophysicists. It is inconceivable that the earth was born much later the sun. Hence gravitational contraction cannot be the source of solar energy.

Thus, the sun has been radiating energy at the rate of 4×10^{26} J $\mathrm{s^{-1}}$ for billions of years. Such an enormous energy could not possibly be due to any chemical reaction, such as burning of coal or oil. Even if the entire mass of the sun is assumed to consist of pure carbon, its complete combustion would yield radiant energy at the above rate for only a few thousand years. We know from Einstein's mass-energy equivalence relation

$$E = mc^2$$

that an enormous amount of energy is released in a nuclear process involving only a small quantity of mass m of the substance, since even for a small m, E can be very large due to a very large magnitude of the factor c^2 ($c = 3 \times 10^8$ m $\mathrm{s^{-1}}$). It was, therefore, postulated that only a nuclear process could be responsible for the production of solar energy. There are two nuclear processes that release energy, namely, nuclear fission and nuclear fusion. The process of fission involves the splitting of heavier unstable elements (such as uranium) into lighter stable elements. This process was ruled out because the abundance of heavier elements in the sun is too small to account for its high rate of emission for so long a period.

It was Hans Bethe of Cornell University, USA, who in 1939 proposed that the process of nuclear fusion could be the only source of solar energy. This process involves a fusion (fusing together) of lighter elements to form slightly heavier elements. The sun has an almost inexhaustible amount of light elements such as hydrogen and helium which together constitute about 98% of the mass of the sun. Bethe suggested a set of nuclear reactions to account for the solar energy and it is called the carbon-nitrogen cycle. Recent experiments have shown that another fusion process called the proton-proton cycle is the most probable source of solar energy. In both these cycles, four hydrogen atoms fuse together to form a helium atom producing two positrons (positive electrons) and releasing energy according to the equation.

$$4\,{}_1^1\mathrm{H} \rightarrow {}_2^4\mathrm{He} + 2e^+ + Q$$

Now the atomic mass of four hydrogen atoms is 4.032 and that of one helium atom is 4.003. Thus, there is a loss of about 0.7 per cent of mass. This is converted into energy in accordance with Einstein's mass-energy relation stated above. Calculations have shown that the energy released in this process is enough to maintain the radiation of energy from the sun for billions of years.

15.6 THE STARS

As mentioned earlier, the sun is a typical star of average size, mass and brightness. There are millions of stars in the universe. Some of them are much more massive and brighter than the sun. In the following we shall discuss distances, radii, brightness and other aspects concerning stellar evolution.

Stellar Distances

Within the solar system the distances can be conveniently expressed in terms of the astronomical unit. Stellar distances, being much larger, are expressed in terms a much bigger unit, the light year. As stated earlier, one light year is equal to a distance of 9.46×10^{15} m. The distances of stars are also measured by the method of triangulation. For very distant stars spectroscopic methods are used. The star nearest to the earth after sun is named Alpha Centauri. It is 4.3 light years away, i.e light starting from the star takes 4.3 years to reach us. The sun is only about 8 light minutes away from us.

Brightness of Stars

When we scan the sky on a clear night some stars look bright and others faint. A star looks bright either because it is intrinsically (really) bright or because it is closer to us. The sun looks the brightest of all stars because it is closest to us, although it is really much fainter than many other stars which appear to be dimmer because of their enormous distances. Thus, what we perceive is only the apparent brightness.

The real brightness of stars is represented through a system of *magnitudes*. This system was started by the Greek astronomer Hipparchus about 2000 years ago. The brightest visible stars were called the first magnitude stars. Stars which are half as bright were called the second magnitude stars and so on up to the sixth magnitude stars which were the faintest stars in the sky. With the use of good telescopes, it is now possible to see very faint stars. Consequently, the magnitude scale of Hipparchus has now been extended far beyond six. Photometric measurements show that the first magnitude stars of Hipparchus are actually about 100 times brighter than his sixth magnitude stars. The new magnitude scale is based on these measurements; it is defined such that a magnitude difference of 5 corresponds exactly to a factor of 100 in the amount of light energy reaching the earth. A difference in magnitude by 1, therefore, corresponds to a difference in the ratio of luminosity (or brightness) by $(100)^{1/5}$ which is about 2.5. This means that, if two luminous objects have luminosities L_1 and L_2, their magnitudes m_1 and m_2 are related as

$$\frac{L_1}{L_2} = 100^{(m_2-m_1)/5} \tag{15.16}$$

Taking the logarithm of both sides, we get

$$\log\left(\frac{L_1}{L_2}\right) = \frac{(m_2-m_1)}{5}\cdot\log 100 = \frac{1}{2.5}(m_2-m_1)$$

$$\text{or}\qquad m_2 - m_1 = 2.5\log\left(\frac{L_1}{L_2}\right) = -2.5\log\left(\frac{L_2}{L_1}\right) \tag{15.17}$$

This equation is used to define the magnitude m of the star. For this purpose, one of the luminous objects, is taken as a standard and its magnitude is taken to be $m = 0$ corresponding to its luminosity L_0. In terms of this standard, the magnitude m of a star of luminosity L is defined as

$$m = -2.5\log\left(\frac{L}{L_0}\right) \tag{15.18}$$

From this definition it follows that stars which are brighter than those of zero magnitude have negative m values. Thus, a magnitude of -5 corresponds to brightness 100 times greater than that of the zero magnitude star. The value of m for the sun is about -26.5 while that for venus is about -4.

EXAMPLE 15.7 Venus is about 10^4 times brighter than the dimmest visible star. If the magnitude of the dimmest star is $+6$, what is the magnitude of Venus?

Solution: $m_1 = +6 \quad \frac{L_2}{L_1} = 10^4 \quad m_2 = ?$

Using Eq. (15.17), we have

$$\begin{aligned} m_2 &= m_1 - 2.5\log\left(\frac{L_2}{L_1}\right) \\ &= 6 - 2.5\log(10^4) \\ &= 6 - 2.5\times 4 = -4 \end{aligned}$$

Astronomers have classified and mapped a large number of stars. Calculations show that their absolute brightness varies in the range of about a million times to one millionth of the brightness of the sun.

Stellar Spectra

The spectrum of the sun has been extensively studied by spectroscopists and astronomers. The solar spectrum is a continuous spectrum on which dark lines (called Fraunhoffer lines) are superimposed as described in Chapter 11. Just like the sun, the stars also show a continuous spectrum on which dark absorption lines are superimposed.

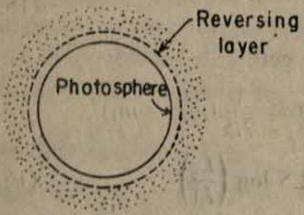


Fig. 15.7 *Formation of continuous and absorption stellar spectra*

The continuous spectrum is produced by the surface layers of the star which form what is called the photosphere of the star (Fig. 15.7).

Like a black body, the photosphere of a star emits radiations of all wavelengths. When the continuous radiation of the photosphere passes through the topmost relatively cooler layer (called the *reversing layer*) of the star, the radiations of certain wavelengths are selectively absorbed by this layer. The dark absorption lines in the continuous spectrum correspond to these absorptions. The wavelengths corresponding to the dark lines are characteristic of the substances present in the reversing layer. Thus, we can identify the elements present in the stars by noting the positions of the absorption lines in their spectra.

The spectra of all stars are identical. The spectrum of a star is related to its colour. Stellar spectra have been divided into seven principal classes, namely, O, B, A, F, G, K and M. The O and B type stars are blue and their spectra have helium lines. The A type stars are white and their spectra show hydrogen lines predominantly. The F and G type stars are green and yellow: their spectra show lines of ionised neutral metals. The K type stars are orange and M type stars are red in colour. Their spectra show absorption bands that indicate the presence of molecules of titanium oxide in them.

The colour of a star is related to its surface temperature. This can be understood from Wien's displacement law discussed in Class XI. According to this law

$$\lambda_{\max} T = \text{Constant}$$

where T is the temperature and λ is the wavelength at which the intensity of radiation is maximum. Now the visible radiations are VIBGYOR (violet, indigo, blue, green, yellow, orange and red) with violet having the shortest wavelength and red the longest. From the displacement law, it is clear that very hot stars will predominantly emit radiation of short wavelengths. They, therefore, look indigo or blue. Less hot stars look green or yellow: whereas cooler stars look orange or red. Notice the colours of an iron ball when it is heated. When it is cold, it looks black. When heated, it first becomes red hot and on further heating the colour changes to yellow, then to white or blue.

The surface temperature of a star can be estimated by measuring the wavelength at which the intensity of the emitted radiation is maximum and then using the displacement law. After knowing the temperature, the true chemical composition of the atmosphere of a star can be determined. H.N. Russell has done pioneering work in this field. He found that all stars have essentially the same composition as that of the sun, i.e. mostly hydrogen and helium with very little traces of heavier elements.

Stellar Radii

Stars are so far away from the earth that even the largest stars when viewed with the most powerful telescopes appear as mere point images. The angular diameters of these images are so small that the method described in Sec. 15.4 cannot be used to measure them and we have to resort to indirect methods.

We can use the Stefan-Boltzmann law of heat radiations to estimate the radius of a star. Assuming that the star emits radiation like a black body, then according to the Stefan-Boltzmann law, the amount of radiant energy emitted per second by a unit area of the surface of the star is given by σT^4 where σ is Stefan's constant and T the surface temperature of the star. If r is the radius of the star, then the total energy radiated per second by its entire surface $(4\pi r^2)$ is given by

$$E = (\sigma T^4)(4\pi r^2) = 4\pi\sigma r^2 T^4 \tag{15.19}$$

This equation will give the radius r of the star if its temperature T and intensity E are known. How these measurements are made has already been discussed above. For most stars, the radii range from about 20 solar radii to about a tenth of the solar radius.

Giants and Super Giants There are a few exceptions. The stars named Capella and Betelgeuse have radii equal to 50 and 220 times the radius of the sun. On account of their large size, the stars are called giants and super giants.

Dwarfs and White Dwarfs The most common stars like the sun are called dwarfs by comparison with giants. On the other extreme, very small stars like Sirius whose diameters are a fiftieth of that of the sun or less are called white dwarfs.

Stellar Masses

Many stars are binary systems. A binary system consists of a pair of stars revolving around their common centre of mass. The masses of the stars in a binary system are determined by the application of Kepler's law of periods as discussed in Sec. 15.4.

Let us suppose that M_1 and M_2 are the masses of the two stars in a binary system revolving about their common centre of mass C in orbits of radii r_1 and r_2 (see Fig. 15.4). Then we have from Eq. (15.6).

$$M_1 + M_2 = \frac{4\pi^2}{G} \cdot \frac{r^3}{T^2}$$

where $r = r_1 + r_2$, T their common period of revolution and G is the universal gravitation constant. For convenience let us take Sun's mass as the unit of mass, the astronomical unit (AU) as the unit of distance and the year as the unit of time. It is easy to see that, in terms of these units, $G = 4\pi^2$, so

$$M_1 + M_2 = \frac{r^3}{T^2}$$

where M_1 and M_2 are in units of solar mass, r in AU and T in years.

Let us consider a typical example. Sirius is a binary; its two components are separated by a distance $r = r_1 + r_2 = 20.3$ AU. Observations show that they revolve about their common centre of mass in 49.9 years. Thus $T = 49.9$ years. Substituting these values in the above equation, we have

$$M_1 + M_2 = 3.35M$$

where M is the mass of the sun. By locating the centre of mass, it is found that $r_2/r_1 = 2.35$. Hence, we get

$$M_1 = 2.35M$$

$$M_2 = 1.00M$$

Similar measurements have been made for about 100 binary systems. The masses of stars range from about $40M$ for the hottest O type stars to about $0.1M$ for the coolest M type stars.

Interiors of Stars

If we divide the mass of a star by its volume, we get its mean density. The mean densities of dwarfs range from 10 kg m^{-3} (O stars) to 5×10^4 kg m^{-3} (M stars). As in the case of the sun, the temperature and density increase as we go deeper into the interior of a star. The central densities of dwarfs range from a few thousand kg m^{-3} (O type) to about 10^4 kg m^{-3} (M type). The central temperature is maximum for O and B type stars (30×10^6 K) and minimum for M type stars (10×10^6 K).

Stellar Evolution

Like everything else in the universe, a star is born at a certain point of time, it lives for a certain length of time and eventually dies.

Birth of a Star

Interstellar space is filled with particles of dust and gas. The weak gravitational attraction between them draws them together to form a cloud. The life history of a star, according to widely accepted theory, begins when the cloud of dust and gas becomes large enough to contract under the effect of its own gravitational force. Due to compression, the cloud heats up. It begins to radiate (or lose) energy, which helps the cloud to contract further. Detailed calculations show that for a cloud to contract, its original mass must be at least one thousand solar masses. At a certain stage, the cloud explodes into smaller fragments, each of which continues to contract under its own weight. When the fragments become hot enough, they begin to radiate light from their surface. Each fragment becomes a self-luminous star. Thus, a cluster of stars is born.

Death of a Star

Each star of a cluster continues to contract until its core attains a temperature of about 10 million degrees. At these temperatures the nuclear fusion process, discussed in Sec. 15.5, starts. In this process, hydrogen is converted into helium and energy is released. This energy keeps the star shining for millions of years. The dwarfs, which form more than 90% of the stellar population, are in this stage of evolution. This stage lasts until all the hydrogen fuel in the core of the star is exhausted. Our sun has already spent about 5 billion years of its life and it will take another 5 billion years for its hydrogen fuel to be exhausted.

When the hydrogen in the core of the star is exhausted, the star enters its old age. The nuclear process stops. The core of the star starts contracting again. This raises the temperature of the core which heats up the outer layers of the star, causing them to expand. Due to expansion, the outer layers cool. At this stage the size of the star becomes very big and its temperature falls, making it look red. It becomes what is called a *red giant*. With continued contraction of the core and the resulting rise in its temperature, helium in the star may become the fuel for thermonuclear reactions to form still heavier elements in the star. Eventually with the depletion of the fuel elements, a violent explosion called *nova* or *super nova* may occur in which a large portion of the star's envelope is thrown back into interstellar space. This is the death of the star.

The core that remains behind may end up as one of the following three types of stellar corpses (or dead bodies).

1. If the original mass of the star was less than $2M$ (where M is the mass of the sun), we get a dense white dwarf. As there is no nuclear fuel left in the dwarf, it changes colour from white to yellow to red as it cools. Finally it becomes black, and is lost to sight for ever.
2. If the original mass of the star was between $2M$ and $5M$, the recoil of the super nova explosion compresses the core to nuclear densities ($\sim 10^{17}$ kg m^{-3}) giving rise to what is called a *neutron star*. Its mass is less than $2M$ and its radius is about 10 km. Neutron stars have very high magnetic fields ($\sim 10^8$ T).
3. If the original mass of the star was greater than $5M$, the recoil of the super nova explosion is so violent that the core continues to collapse indefinitely giving rise to what is called a *black hole*. A black hole has a very small size but enormous surface gravity, so much so that even a light signal (photon) cannot escape from its surface. On the other hand, any particle or photon that approaches it is attracted and swallowed. That is why such a body is called a black hole.

Pulsars and Quasars

Pulsars Although the existence of neutron stars was postulated over thirty years ago, it was only recently that their existence was confirmed observationally with the accidental discovery of what are called *pulsars*. The first pulsar (CP 1919) was discovered in 1967 by astronomers at Cambridge University in England. The radio receiver at their observatory received a weak radio signal which repeated with great regularity. It sent pulses at a regular interval of precisely 1.33730115 s. Later about a dozen such objects were discovered. It is generally believed that the pulsars are high density neutron stars having a very rapid rotation and a large magnetic field as stated before.

Quasars

Another very spectacular discovery in astronomy during the last twenty-five years is that of quasars—a short form for quasi-stellar radio sources. The first quasar discovered was named 3C 273. A quasar is an object that emits enormous power, of the order of 10^{53} W (or J s^{-1}) which is equivalent to the energy of a million Suns. Spectroscopic observations show that they move with very high speeds approaching 0.9 times the speed of light. It is believed that such an enormous power output of a quasar may be due to gravitational collapse. They appear as faint point images. About 150 quasars are known.

15.7 THE MILKY WAY

The universe consists of a very large number (about 10^9) of galaxies. Each galaxy, in turn, has a large number of stars (about 10^{11}) clustered in it. The name galaxy has been derived from the word 'gala' which means milk. Our own galaxy is called the *Milky Way*.

Size and Shape

The Milky Way galaxy is in the form of a disc which is thicker near the centre and thinner towards the edges. Figure 15.8 shows a section of this galaxy in the plane of the paper. The diameter of the disc is about 10^5 light years. The thickness of the disc is about 5000 light years at the centre and diminishes near the edges. The sun, which belongs to this galaxy, is situated in its central plane at a distance of about 2.7×10^4 light years from the centre of the galaxy (called the *galactic centre*). When seen through a telescope, it gives a cloudy appearance due to the presence of a very large number of stars in it.

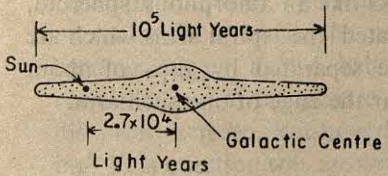


Fig. 15.8 *Edge-on view of the Milky Way (schematic illustration)*

Interstellar Matter

On a clear night, we see several dark regions in the band of the Milky Way. The regions look dark due to the presence of dust clouds which hide from our view the stars lying behind them. The space between stars in the Milky Way is filled with dust and gas. Sometime the dust gets illuminated by a hot star and it begins to shine by light reflected from it giving rise to what is called a *nebula*. Orion is a bright nebula. About 90% of the interstellar matter is composed of hydrogen.

Stellar Clusters

The smaller assemblies of stars in a galaxy are known as clusters. There are two types of star clusters, namely, the *galactic clusters* and the *globular clusters*. Galactic clusters are groups of 100 to 1000 stars held in position by mutual gravitational forces. The cluster moves as a whole around the galaxy. *Globular clusters* are groups of about 10,000 stars packed in a relatively small region. About 1000 globular clusters are known. All of them are more than 20,000 light years away from the sun.

Structure

The disc of the Milky Way has a spiral structure. Ordinary optical telescopes cannot be used to study the detailed structure because the field of view is blocked by interstellar dust clouds. But radiowaves can penetrate the clouds of dust. Tharefore, radio telescopes have recently been used to study the detailed structure. These studies have revealed that the central part of our galaxy looks like an amorphous spheroid. Outside the central nucleus are situated the spiral arms which are about 1,200 light years wide and are separated by gaps of about 500 light years. The sun is situated near the edge of one such arm.

Rotation

A remarkable feature of the Milky Way is that it rotates about an axis passing through its centre. This rotation is not like that of a rigid rotator. Each star of the galaxy moves in an orbit with a speed that is determined by the gravitational attraction of the stars enclosed within its orbit. The sun is one of about 150 billion stars in the Milky Way. All the stars revolve round the galactic centre. The sun, moving at a speed of 250 km s^{-1}, revolves about the centre in about 250 million year.

Mass

The mass of the Milky Way can be estimated as follows: If r is the distance of the sun from the galactic centre and v its speed round its orbit, the centripetal acceleration is v^2/r. If M is the mass of the Milky Way, the gravitational acceleration of the sun will be GM/r^2. Equating the two we get

$$M = \frac{v^2 r}{G}$$

Substituting for $v = 2.5 \times 10^5 \text{ ms}^{-1}$, $r = 2.7 \times 10^4$ light years $= 2.7 \times 10^4 \times 9.46 \times 10^{15} \simeq 3.5 \times 10^{20}$ m and] $G = 6.67 \times 10^{-11} \text{ Nm}^2 \text{ kg}^{-2}$, we get $M \simeq 3 \times 10^{41}$ kg, which is about 150 billion solar masses. From this, one may conclude the part of our Milky Way within the orbit of the sun consists of around 150 billion stars.

15.8 COSMOLOGICAL THEORIES OF THE UNIVERSE

The Expanding Universe

There are millions of galaxies in the universe. It has been established by observations that these galaxies are all receding away from us and from one another at fantastic speeds. Thus, the universe is expanding. Let us take the analogy of a balloon with several ink dots on it. When the balloon is inflated, each ink dot moves away from the other during expansion. The distance between any two ink dots keeps on increasing as the expansion of the balloon is continued. However, there are two points of difference between the expanding universe and the expanding balloon. The balloon expands about its centre; but there is no centre about which the universe expands. Secondly, the ink dot becomes bigger as the balloon is inflated, but in the expanding universe the sizes of the galaxies do not change; only the inter galactic distances increase. But how do we know that the universe is expanding? This question was answered by a well-known astronomer named Hubble. It is a matter of common experience that the frequency of the whistle of an incoming train appears to increase while that of a departing train -appears to decrease. This is called the Doppler effect. This effect occurs in the case of light also. Thus the radiation from an approaching source of light should have a higher frequency (therefore, a shorter wavelength) and that from a receding source of light should have a lower frequency (or longer wavelength) than when the source is not in motion. Thus a receding source of light should show a shift towards the longer wavelength region (i.e. the red region) in the lines of its spectrum. This is known as the *red shift*. In fact the speed of recession is estimated from the amount of red shift of the spectral lines. Hubble observed that the light from distant galaxies shows a red shift. Thus it follows that the galaxies are receding. He found that the speed of recession (v) is related to the distance (r) of the galaxy by the relation

$$v = Hr$$

This is known as Hubble's law and the constant H whose value is 17 kms^{-1} per million light years is called *Hubble's constant*. Thus, the higher the distance of a galaxy, the higher is the speed of its recession. Knowing the speed of recession from the red shift, one can determine the distance of a galaxy from Hubble's equation.

A natural question to ask is: What causes the expansion of the universe? This brings us to the question of the origin of the universe. The following three theories have been proposed to account for the origin and evolution of the universe.

The Big Bang Theory

This theory proposes that, in the remote past, all the matter now present in the universe was concentrated in a single extremely dense and hot ($\sim 10^{12}$ K) fireball. Then an explosion somehow occurred about 20 billion years ago and the ball broke into large number of segments which flew off in all directions in the form of galaxies. According to Hubble's law, the speed of recession of a galaxy becomes equal to the speed of light at a distance of about 20 billion light years. No light from a galaxy at such a distance would ever reach us so that this distance becomes the boundary of the observable universe. Due to continuous recession, more and more galaxies would be lost from our view, as they approach this distance.

The Pulsating Theory

Some astronomers believe that the expansion of the universe cannot continue and is stopped by the gravitational pull and the universe begins to contract again. After it has contracted to a certain critical size, an explosion again occurs. According to this theory the expansion and contraction repeat after every 8 billion years. Thus, the outer boundary of the universe moves in and out (i.e. pulsates or oscillates) at regular intervals.

The Steady State Theory

This theory was proposed by astrophysicists named Bondi, Gold and

Hoyle. According to this theory, the galaxies which have receded beyond the observable part of the universe are replaced by new galaxies which are continuously being formed out of the interstellar matter. This theory, therefore, suggests that the universe has always appeared as it does today and the rate of expansion has been the same in the past and will remain the same in future. Some kind of a steady state has been reached so that the total number of galaxies in the universe remains constant.

There is not enough observational evidence to decide which of these theories is more plausible than the other. A fairly large number of cosmologists believe the Big Bang theory to be the most probable. The evidence in favour of this theory came in the sixties. The 'fireball' which was extremely hot in the remote past must be steadily cooling due to expansion. Assuming that the radiation emitted from the fireball during the cooling process has a black-body spectrum, the wavelength at which maximum intensity occurs is related to temperature T by Wien's law: $\lambda_{\max} T = \text{constant}$. The temperature T of this radiation spread out over the entire universe can be shown to be about 3 K. In 1965 Arno Penzias and Robert Wilson succeeded in detecting such an all-pervading radiation. This is the basis of the wide-spread acceptance of the Big Bang theory of the evolution of the universe.

We now have a fairly good idea of the large scale structure and evolution of the universe. But a large number of questions still remain unanswered. For instance, we do not know how galaxies were formed,

TABLE 15.1 Physical Properties of Planets of the Solar System
1AU = 1.496×10^{11} metre, $R_E = 6.378 \times 10^6$ metre, $M_g = 5.977 \times 20^{24}$ kg, $g_E = 9.8$ m s^{-2} (subscript E refers to Earth) Abbreviations: d = day, h = hour, m = minute, 1 year = 365.257 days.

Planet	Period of revolution in years	Semi-major axis of orbit (AU)	Radius, ($\times R_E$)	Rotation period	Mass, ($\times M_E$)	Mean density (kg m^{-3})
Mercury	0.241	0.387	0.38	58.6d	0.056	5400
Venus	0.615	0.723	0.96	243d (retrograde)	0.815	5100
Earth	1.000	1.000	1.00	23h 86 m	1.000	5520
Mars	1.881	1.524	0.53	24h 27m	0.107	3970
Jupiter	11.864	5.203	11.23	9h 50 m	317.9	1330
Saturn	29.46	9.540	9.41	10h 14m	95.0	700
Uranus	84.0	19.18	3.98	10h 49m (retrograde)	14.6	1330
Neptune	164.1	30.07	3.88	15h	17.2	1660
Pluto	247	39.44	0.50	6.39d	0.11	4900

why the universe is so homogeneous and how exactly the universe was created in the first place. Hopefully these and other questions will be answered one day with improvements in observational and theoretical cosmology.

SUMMARY

The branch of science that deals with the study of the universe is called astronomy. Astronomy is an observational science. The devices that an astronomer uses for this study include optical telescopes, radio telescopes, spectrograph, photographic plates, space probes, etc.

The main constituents of the universe are the solar system, stars and galaxies.

The solar system consists of the sun at the centre, nine planets (including the earth) revolving around the sun with 32 natural satellites that revolve around the planets. In addition, there are more than 2000 asteroids and several hundred comets.

The average distance between the sun and the earth is called astronomical unit (AU) IAU $= 1.496 \times 10^{11}$ m.

The distances of the objects in the solar system can be measured by the method of triangulation and by the radar echo method. The mass of a planet can be obtained from the knowledge of its period of revolution and its distance from the sun. The composition of the atmosphere of a planet and its surface temperature are determined from the study of the spectrum of the radiation reflected from it.

The sun is a typical star. It is one of about 100 billion stars in the Milky Way. The diameter of the sun is about 7×10^8 m and its surface temperature is about 6000 K. The process of thermonuclear fusion is responsible for the enormous energy radiated by the sun. Stellar distances are expressed in terms of light years. 1 light year $= 9.46 \times 10^{15}$ m.

A cluster of a large number of stars is called a galaxy. There are more than 10^9 galaxies in the universe. Our galaxy is called the Milky Way. Galaxies can be spiral, elliptical and irregular. Most galaxies are spiral like our own galaxy.

The branch of astronomy that deals with the origin, structure and evolution of the universe is called cosmology. The measurements of the red shift of spectral lines suggest that the universe is expanding. Three theories have been proposed to account for the origin and evolution of the universe. They are the big band theory, the steady state theory and the pulsating theory. The red shift observed for quasars and the other observations tend to favour the big band theory.

EXERCISES

Short-Answer Questions

1. Name the planets in the solar system.
2. What is a light year?
3. Name the galaxy to which we belong.
4. What is the surface temperature of the sun?
5. What is a galaxy?

Long-Answer Questions

1. Distinguish between a planet and a star.
2. Name the constituents of the solar system. How will you determine the: (i) distance, (ii) mass, (iii) size, (iv) period of rotation, and (v) surface temperature of a planet?
3. Explain clearly why some planets have atmospheres while others do not.
4. State the important physical properties of the sun. How is the surface temperature of the sun estimated?
5. Discuss the possible source of solar energy.
6. How is the surface temperature of a star computed?
7. Narrate the complete life history of a star.
8. How can we determine the composition of the outer layers of a star by studying its spectrum?
9. Describe the structure and contents of the Milky Way.
10. What is a galaxy? Distinguish between a normal galaxy and a radio galaxy.
11. How can we find out that the universe is expanding? Briefly describe the three theories of the evolution of the universe.
12. Write short notes on the following: (i) electromagnetic radiations, (ii) comets, (iii) asteroids, (iv) meteor, (v) meteorites, (vi) solar constant, (vii) photosphere, (viii) thermonuclear fusion, (ix) sunspots, (x) solar flares, (xi) prominences, (xii) absorption spectrum of a star, (xiii) red giants, (xiv) white dwarfs, (xv) nova and super nova, (xvi) neutron star, (xvii) black hole; (xviii) pulsars, (xix) quasars, (xx) galactic and globular clusters, (xxi) spiral galaxies, (xxii) radio galaxies, (xxiii) the red shift, and (xxiv) Hubble's law.

C. Multiple Choice Questions

Choose the correct answer from the given alternatives.

1. Which planet is closest to the sun?

   (a) Earth (b) Mercury
   (c) Mars (d) Jupiter

2. Which planet is the farthest from the sun?

   (a) Saturn (b) Uranus
   (c) Neptune (d) Pluto

3. The group of small pieces of rock revolving round the sun between the orbits of Mars and Jupiter are called
   (a) Meteors (b) Comets
   (c) Meteorites (d) Asteroids
4. The density of matter in a neutron star is of the order of
   (a) 10^3 kgm^{-3} (b) 10^7 kgm^{-3}
   (c) 10^{17} kgm^{-3} (d) 10^{21} kg m^{-3}
5. A light signal (photon) cannot escape from the surface of a
   (a) black hole (b) neutron star
   (c) white dwarf (d) red giant

D. Numerical Problems

1. Phobos is a satellite of Mars. The period of its revolution around Mars is 0.319 days and its mean distance from Mars is 9.519×10^6 m. Assuming that the mass of Phobos is negligible compared to that of Mars, calculate the mass of Mars. Take $G = 6.668 \times 10^{-11}$ N m^2 kg^{-2}.
2. The period of revolution of the moon around the earth is 27.32 days and its mean distance from the earth is 384,400 km. If the centre of mass of the earth-moon system is located at a distance of 4.75×10^6 m from the earth's centre, calculate the mass of each. Given $G = 6.668 \times 10^{-11}$ N m^2 kg^{-2}.
3. The period of revolution of Mars around the sun is 1.881 year and its mean distance from the sun is 1.524 AU. Calculate the mass of the sun (1 AU = 1.496×10^{11} m).
4. Calculate the value of the solar constant on Mars taking its distance from the sun as 1.524 AU. Also calculate the temperature of a black body on Mars. Solar constant on Earth = 1.388 10^3 W × m^{-2}, Stefan's constant = 5.669×10^{-8} W m^{-2} K^{-4}.
5. The total power emitted by the sun is 3.9×10^{26} W. The mean distance of the sun from the earth is 1.496×10^{11} m. Calculate the value of the solar constant.
6. The moon subtends an angle of 57° at the base line equal to the radius of the earth. What is the distance of the moon from the earth? The radius of the earth = 6.378×10^6 m.
7. Compute the surface temperature of the sun from the following data:
   Distance of Mars from the sun = 1.524 AU
   Solar constant of Mars = 5.976×10^2 W m^{-2}
   Stefan's constant = 5.735×10^{-5} W m^{-2} K^{-4}
   Radius of the sun = 6.928×10^8 m
8. Assuming that the faintest star visible to the naked eye has a magnitude 6, compare its brightness with that of planet Venus whose magnitude is 4.
9. The diameter of the planet Jupiter is 3580 km. What is its angular diameter when it is a distance of 8.25×10^8 km from the earth?
10. Suppose the sun suddenly shrank to half its present size. Find the gravitational potential energy released in this process.

APPENDIX A

LINE, SURFACE AND VOLUME INTEGRALS

It is worthwhile to recall the definition of an ordinary single integral which you are already familiar with. Suppose the function $f(x)$ of a single real variable x is given (or known) for all values of x in an interval $x = \mathrm{A}$ to $x = \mathrm{B}$, where A and B are the end-points on a linear interval. Then the integral

$$I = \int_{\mathrm{A}}^{\mathrm{B}} f(x)\, dx \tag{A.1}$$

of the function $f(x)$ from $x = \mathrm{A}$ to $x = \mathrm{B}$ is defined as follows. We divide the interval from A to B into a very large number N of subintervals of very small lengths, as shown in Fig. A.1, where the end-point A is labelled P_0 and the end-point B is labelled P_N and P_1, P_2, P_3, . . . P_{N-1} are the intermediate points. The lengths of the subintervals are Δx_1, Δx_2, . . . Δx_{N-1}, Δx_N as shown in the figure.

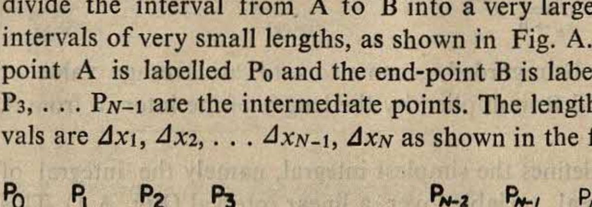


Fig. A.1 *Definition of a simple integral of a single real variable*

We could, of course choose the intermediate points P_1, P_2, . . . , P_{N-1} at equal distances such that all the subintervals have the same length; but it is not essential since ultimately we take the limit $\Delta x \to 0$. Then we choose one point in each subinterval, say x_1, x_2, x_3 . . . x_{N-1}, x_N; x_1 is a point in the first subinterval, x_2 in the second, and so on. It does not matter where in a subinterval we choose this point, because if the subinterval is small enough (strictly speaking tending to zero).

the value of the function $f(x)$ will be sensibly constant over a subinterval.

Next we multiply the value of the function f at a point x in a subinterval by the length of that subinterval and add all such products for all the subintervals to get the sum

$$
\begin{aligned}
I_N &= f(x_1)\Delta x_1 + f(x_2)\Delta x_2 + \ldots + f(x_N)\Delta x_N \\
&= \sum_{n=1}^{n=N} f(x_n)\Delta x_n \qquad \text{(A.2)}
\end{aligned}
$$

where $f(x_n)$ is the value of the function at any point in the nth subinterval whose length is Δx_n and n takes on values $1, 2, \ldots, N$; $n = 1$ for the first subinterval, $n = 2$ for the second, . . . and $n = N$ for the Nth subinterval.

We then take the limit $N \to \infty$ and $\Delta x_n \to 0$, i.e. the number of the subintervals becomes larger and larger and the size of the subinterval becomes smaller and smaller. In this limit, the sum I_N approaches a definite value which is called the integral I of $f(x)$ over the limits A and B (x_0 and x_N), i.e.

$$
I = \underset{\substack{N\to\infty \\ \Delta x_n \to 0}}{\text{limit}} \sum_{n=1}^{n=N} f(x_n)\Delta x_n = \int_{A(=x_0)}^{B(=x_N)} f(x)\, dx \qquad \text{(A.3)}
$$

The integral I may be regarded as some sort of an average value of $f(x)$ between A and B multiplied by the length of the interval from A to B.

Equation (A.3) defines the simplest integral, namely the integral of a function of a real variable over a linear interval (Fig. A.1). This mathematical definition of an integral can be generalized to the following more complicated situations:

(a) The region of integration, instead of being a linear interval, could be (i) a curve in a plane or in space (one-dimensional case), (ii) a surface (or area) in a plane or in space (two-dimensional case), or (iii) a volume in space (three-dimensional case).

(b) The function to be integrated (i.e. the integrand) could be a vector function instead of a scalar function.

The three cases (i), (ii) and (iii) lead us to the concepts of line, surface and volume integrals of scalar and vector functions. In each case the procedure to calculate the integral is the same—we divide the curve, surface or volume into a very large number of elements of very small sizes, multiply the value of the function at any point in an element by the size of that element, add up all such products for all the elements and see if the sum approaches a definite value as the number of elements tends to infinity and the size of each element tends to zero. That value is defined as the line, surface or volume integral.

LINE INTEGRAL

Consider two points A and B in, say the x-y plane and let L be a continuous curve between points A and B in the plane. Let $\mathbf{r}(x, y)$ be the position vector of a point in the curve. Let $f(\mathbf{r})$ be a scalar function whose value is given for all points $\mathbf{r}(x, y)$. Divide the curve L into N small segments (not necessarily of the same length), as shown in Fig. A.2 where the end-point A is labelled P_0, the end-point B is labelled P_N and $P_1, P_2, P_3, \ldots, P_{N-1}$ are the intermediate points.

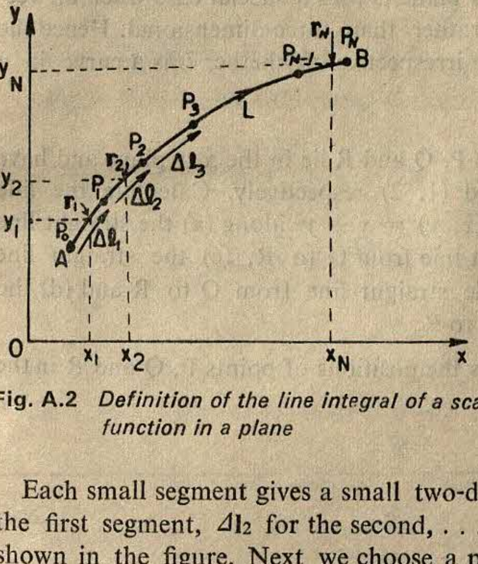


Fig. A.2 *Definition of the line integral of a scalar function in a plane*

Each small segment gives a small two-dimensional vector: $\Delta \mathbf{l}_1$ for the first segment, $\Delta \mathbf{l}_2$ for the second, . . . $\Delta \mathbf{l}_N$ for the last segment as shown in the figure. Next we choose a point $\mathbf{r}_1$ anywhere in the first segment, a point $\mathbf{r}_2$ anywhere in the second, . . . and a point $\mathbf{r}_N$ in the last segment. Here $\mathbf{r}_1(x_1, y_1)$, $\mathbf{r}_2(x_2, y_2)$, . . . , $\mathbf{r}_N(x_N, y_N)$ are the position vectors of the points chosen. Then we multiply the value of the function at that point in an element by the magnitude of the length vector of that element and construct the sum of all such contributions from all the segments. We get

$$
\begin{aligned}
I_N &= f(\mathbf{r}_1)\Delta l_1 + f(\mathbf{r}_2)\Delta l_2 + \ldots + f(\mathbf{r}_N)\Delta l_N \\
&= \sum_{n=1}^{n=N} f(\mathbf{r}_n)\Delta l_n
\end{aligned}
$$

Finally we take the limit $N \to \infty$ and $\Delta l \to 0$. If the sum I_N tends to a definite value in this limit, then that value gives the line integral of f along the curve L, i.e.

$$I_L = \int_{\substack{A \\ \text{along } L}}^{B} f(\mathbf{r})\,\mathrm{d}l \tag{A.4}$$

This is the definition of the line integral of a scalar function in two dimensions. The generalization to three dimensions is straightforward. If L is a curve in space from points A to B, the position vector $\mathbf{r}$ is a three-dimensional vector $\mathbf{r}(x, y, z)$ and hence all the small vectors $\Delta\mathbf{l}_1$, $\Delta\mathbf{l}_2$, . . . are also three-dimensional vectors and the points $\mathbf{r}_1, \mathbf{r}_2, \ldots \mathbf{r}_N$ are points in three-dimensional space and lie on the curve L. The above case of L being in a plane is thus a special case when all vectors are two-dimensional rather than three-dimensional. Hence the definition (A.4) is applicable irrespective of whether L is a curve in a plane or in space.

EXAMPLE A.1 Three points P, Q and R lie in the x–y plane and have coordinates (0, 0), (1, 0) and (1, 2) respectively. Calculate the line integral of the function $f(x, y) = x + y^2$ along (a) the straight line from P to Q, (b) the straight line from Q to R, (c) the straight line from P to Q and then the straight line from Q to R and (d) the straight line from P directly to R.

Solution: Figure A.3 shows the positions of points P, Q and R in the x–y plane.

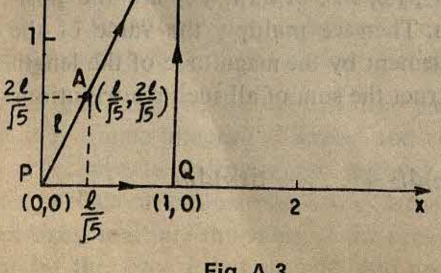


Fig. A.3

(a) The line integral to be evaluated is

$$I_{P\to Q} = \int_P^Q f(x, y)\,dl = \int_P^Q (x + y^2)\,dl,$$

along the straight line from P to Q.
Since $y = 0$ along this line, we have

$$I_{P\to Q} = \int_P^Q x\,dx = \int_{x=0}^{x=1} x\,dx = \frac{x^2}{2}\Big|_0^1 = \frac{1}{2}$$

(b) The line integral to be evaluated is

$$I_{Q\to R} = \int_Q^R f(x, y)\,dl = \int_Q^R (x + y^2)\,dl,$$

along the straight line from Q to R.

Since $x = 1$ along this line, we have

$$I_{Q\to R} = \int_Q^R (1 + y^2)\,dy = \int_{y=0}^{y=2} (1 + y^2)\,dy$$

$$= y\Big|_0^2 + \frac{y^3}{3}\Big|_0^2 = 2 + \frac{8}{3}$$

$$= \frac{14}{3}$$

(c) The integral to be evaluated is

$$I_{P\to Q, Q\to R} = \int_P^Q f(x, y)\,dl + \int_Q^R f(x, y)\,dl$$

$$= I_{P\to Q} + I_{Q\to R}$$

$$= \frac{1}{2} + \frac{14}{3} = 5\frac{1}{6}$$

(d) The integral to be evaluated is

$$I_{P\to R} = \int_P^R f(x, y)\,dl = \int_P^R (x + y^2)\,dl$$

along the straight line from P to R.

To evaluate this integral, we must first express x and y in terms of l. Consider a point A on the line at a distance l from the origin P.

The value of l is 0 at point P and $\sqrt{5}$ at point R. These are the end-points of the line. The distance between them is $\sqrt{5}$. The x and y coordinates of the point A obviously are $l/\sqrt{5}$ and $2l/\sqrt{5}$ respectively. (See Fig. A.3). Putting $x = l/\sqrt{5}$ and $y = 2l/\sqrt{5}$ in the integrand we have

$$I_{\mathrm{P}\to\mathrm{R}} = \int_{\mathrm{P}}^{\mathrm{R}} (x + y^2)\,\mathrm{d}l$$

$$= \int_{l=0}^{l=\sqrt{5}} \left(\frac{l}{\sqrt{5}} + \frac{4l^2}{5}\right) \mathrm{d}l$$

$$= \frac{l^2}{2\sqrt{5}}\Bigg|_0^{\sqrt{5}} + \frac{4}{5}\cdot\frac{l^3}{3}\Bigg|_0^{\sqrt{5}}$$

$$= \frac{\sqrt{5}}{2} + \frac{4\sqrt{5}}{3} = \frac{11\sqrt{5}}{6}$$

Case (c) and (d) illustrate an important fact about line integrals, namely, that *the value of a line integral between two points, in general depends upon the path of integration.* Notice that in case (c) the path is from P to R via Q whereas in case (d) the path is from P directly to R. The value of the line integral in the two cases is different. Only in special situations is the line integral path-independent. For example, we have seen in chapter 2 that the line integral of the electrostatic field between two points is independent of the path between them but depends only on the end-points of the path.

Generalization to Vector Functions

Our definition (A.4) can be easily generalized to include vector functions also. Suppose we have a vector field $\mathbf{F}(\mathbf{r})$ specified at all points $\mathbf{r}$ in space. $\mathbf{F}(\mathbf{r})$ could be an electrostatic field $\mathbf{E}(\mathbf{r})$ or a magnetic field $\mathbf{B}(\mathbf{r})$ or any other vector field.

Suppose a curve L is given along which the line integral of $\mathbf{F}(\mathbf{r})$ is to be evaluated (Fig. A.4). As before, we divide the curve L into a large number of small vector elements $\Delta\mathbf{l}_1, \Delta\mathbf{l}_2, \ldots, \Delta\mathbf{l}_N$; but now construct the sum of the contributions in a different way by taking the dot products $\mathbf{F}(\mathbf{r}_1)\cdot\Delta\mathbf{l}_1$, $\mathbf{F}(\mathbf{r}_2)\cdot\Delta\mathbf{l}_2 \ldots \mathbf{F}(\mathbf{r}_N)\cdot\Delta\mathbf{l}_N$, where $\mathbf{F}(\mathbf{r}_1)$, $\mathbf{F}(\mathbf{r}_2) \ldots \mathbf{F}(\mathbf{r}_N)$ are the values of the vector field at points in the first, second, ... Nth elements respectively. The sum of the dot products is

$$I_N = \sum_{n=1}^{n=N} \mathbf{F}(\mathbf{r}_n)\cdot\Delta\mathbf{l}_n$$

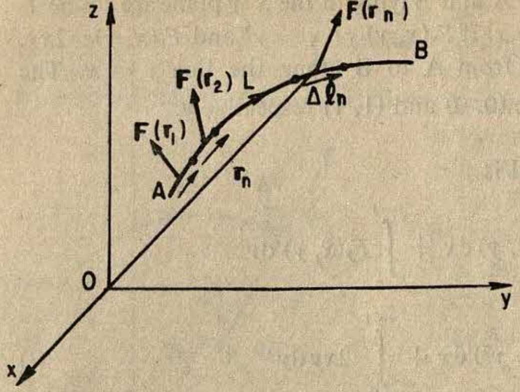


Fig. A.4 *Calculation of the line integral of a vector field in space*

We then take the limit $N \to \infty$ and $\Delta l \to 0$. The sum I_N tends to a value

$$I_L = \int\limits_{\substack{A\\ \text{along } L}}^{B} \mathbf{F}(\mathbf{r})\cdot d\mathbf{l} \tag{A.5}$$

which is the line integral of the vector field $\mathbf{F}(\mathbf{r})$ along the curve L between points A and B. The concept of the line integral of a vector field has been used in chapters 5 and 7.

The dot product in (A.5) means that we take the product of the magnitude of $d\mathbf{l}$ and the tangential component of $\mathbf{F}(\mathbf{r})$ along L. If $\mathbf{F}$ represents a force on a particle (e.g. electrostatic force on a charge in an electrostatic field or gravitational force on a mass in a gravitational field) then the line integral of F over the path described by the particle is the work done by the force in moving the particle from a point A to a point B in the field.

In terms of cartesian components in three-dimensional space (A.5) can be written as

$$\int_{A}^{B} \mathbf{F}(\mathbf{r})\cdot d\mathbf{l} = \int\limits_{x,y,z} \Big\{F_x(x, y, z)\,dx + F_y(x, y, z)\,dy + F_z(x, y, z)\,dz\Big\} \tag{A.6}$$

In the two-dimensional x–y plane (A.6) reduces to

$$\int_{A}^{B} \mathbf{F}(\mathbf{r})\cdot d\mathbf{l} = \int\limits_{x,y} \Big\{F_x(x, y)\,dx + F_y(x, y)\,dy\Big\} \tag{A.7}$$

EXAMPLE A.2 Two points A and B lying in the x–y plane are connected by the straight line $y = x$. If $F_x(x, y) = x - y^2$ and $F_y(x, y) = 2xy$, evaluate the line integral from A to B along the line $y = x$. The coordinates of A and B are (0, 0) and (1, 1) respectively.

Solution: The line integral is

$$I = \int_{x=0}^{x=1} F_x(x, y)\,dx + \int_{y=0}^{y=1} F_y(x, y)\,dy$$

$$= \int_{x=0}^{x=1} (x - y^2)\,dx + \int_{y=0}^{y=1} 2xy\,dy \qquad \text{(i)}$$

Putting $y = x$ in (i) we have

$$= \int_{x=0}^{x=1} (x - x^2)\,dx + \int_{y=0}^{y=1} 2y^2\,dy$$

$$= \frac{x^2}{2}\bigg|_0^1 - \frac{x^3}{3}\bigg|_0^1 + \frac{2y^3}{3}\bigg|_0^1$$

$$= \frac{1}{2} - \frac{1}{3} + \frac{2}{3} = \frac{5}{6}$$

EXAMPLE A.3 What will be the value of the line integral between points A and B in Example A.2 above if the two points are connected by a parabola $y = x^2$?

Solution: Putting $y = x^2$ in Eq (i), we have

$$I = \int_{x=0}^{x=1} (x - x^4)\,dx + \int_{y=0}^{y=1} 2y^{3/2}\,dy$$

$$= \frac{x^2}{2}\bigg|_0^1 - \frac{x^5}{5}\bigg|_0^1 + \frac{4}{5}y^{5/2}\bigg|_0^1$$

$$= \frac{1}{2} - \frac{1}{5} + \frac{4}{5} = \frac{11}{10}$$

EXAMPLE A.4 Suppose a radial force $\mathbf{F}(\mathbf{r}) = C\mathbf{r}$ (where C is a positive constant) acts on a particle at $\mathbf{r}$ in a plane. Find the value of the line integral of $\mathbf{F}(\mathbf{r})$ (a) along an arc of a circle with centre at the origin and (b) along a radial line from A to B at distances r_A and r_B from the origin respectively.

Solution: The line integral is

$$I = \int \mathbf{F}(\mathbf{r})\cdot d\mathbf{l} = \int F(r)\, dl \cos\theta$$

where $F(r)\cos\theta$ is the tangential component of $\mathbf{F}(\mathbf{r})$ and θ is the angle between the vectors **F** and **dl** (Fig. A.5).

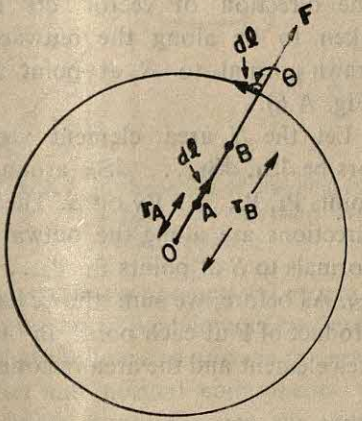


Fig. A.5

(a) $\mathbf{F}(\mathbf{r}) = C\mathbf{r}$ means that, at each point in the plane, the vector **F** is proportional to the position vector of that point **r**. Referring to Fig. A.5, for integration along the arc of a circle, the angle $\theta = 90°$, so that $\cos\theta = \cos 90° = 0$. Hence the integral

$$\int C\mathbf{r}\cdot d\mathbf{l} = 0$$

along the arc of a circle about the origin.

(b) For integration along the radial line from point A to point B, the vectors **F** and **dl** are parallel and the value of the line integral is

$$I = \int_A^B \mathbf{F}(\mathbf{r})\cdot d\mathbf{l} = \int_A^B Cr\, dr \cos 0°$$

$$= C\int_{r_A}^{r_B} r\, dr = \frac{C}{2}\left(r_B^2 - r_A^2\right)$$

SURFACE INTEGRAL

Consider a surface S in three-dimensional space. The surface may be *open*, for example a part of a sphere or a side of a cube; or it may be a closed surface, for example a surface enclosing some volume such as all sides of a cube or the total surface of a sphere or a surface of any arbitrary shape enclosing a volume. To define the surface integral of a vector function **F(r)** over the surface *S*, we divide *S* into a large number (*N*) of surface elements. In order to represent each element of surface geometrically, we recall how we represented a line element in a line integral, We represented each element by a length vector **dl**; the magnitude | **dl** | is the length and the direction was fixed by knowing from which point to which point is the line drawn. Similarly, each elementary area of S around a point P on S is geometrically represented by an area vector **dS**. The direction of vector **dS** is taken to be along the outward drawn normal to S at point P (Fig. A.6).

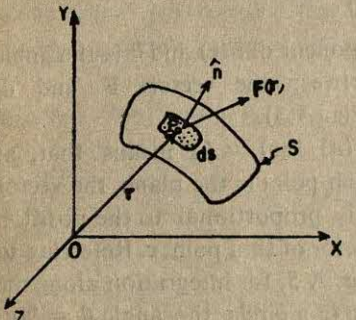


Fig. A.6 *Calculation of a surface integral of a scalar function f*(**r**) *over a volume V*

Let the N area element vectors be $\Delta\mathbf{S}_1, \Delta\mathbf{S}_2, \ldots \Delta\mathbf{S}_N$ around points $P_1, P_2, \ldots, P_N$ on S. Their directions are along the outward normals to S at points $P_1, P_2, \ldots, P_N$. As before, we sum the scalar product of **F** at each point in an area element and the area vector as

$$I_N = \sum_{n=1}^{N} \mathbf{F}(\mathbf{r}_j)\cdot\Delta\mathbf{S}_j$$

In the limit $N \to \infty$ and $\Delta\mathbf{S} \to 0$, we get

$$\mathrm{I}_S = \int_S \mathbf{F}(\mathbf{r})\cdot d\mathbf{S} \qquad \text{(A.6)}$$

$$= \text{(average value of the normal component of } \mathbf{F}) \times \text{(area of } S)$$

While discussing applications of Gauss's theorem in chapter 2, we came across several examples of surface integrals of electric fields. The surface integral of electric field **E**(**r**) over any surface S, open or closed, is called the flux (ϕ_E) of **E** through S. In chapter 7 we have seen that Faraday's law of induction is expressed in terms of the flux (ϕ_B) of the magnetic field through an open surface.

Volume Integral

The third type of integral we use in physics is the volume integral of a scalar function $f(\mathbf{r})$ or a vector field $\mathbf{F}(\mathbf{r})$ over a volume V. For simplicity, we consider only the former. Suppose a scalar function $f(\mathbf{r})$ is specified at all points **r** in volume V. We divide the volume V into a large number (N) of small volume elements $V_1, V_2, \ldots, V_N$ located around points $P_1, P_2, \ldots, P_N$ in volume V. Each volume element is a scalar for it has no direction. As before, the volume integral of $f(\mathbf{r})$ over volume V is given by (see Fig. A.7)

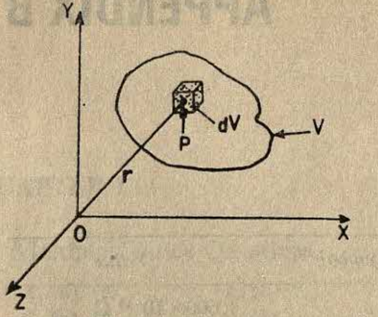


Fig. A.7 *Calculation of volume integral of function f*(r) *over a surface S*

$$I_V = \int_V f(r)\, dV \qquad (A.7)$$

If f is the mass density of some material or the volume density of electric charge, then the volume integral I_V represents the total mass or total charge contained in volume V.

We have seen that the concepts of the line, surface and volume integrals are geometrical generalizations of the ordinary one-variable integrals of functions of one argument. It must be clearly understood that the line and surface integrals of vector fields involve the tangential and normal components respectively of the vector fields. In chapter 2 we have seen the usefulness of these integrals. For example, we saw how the line integral of an electrostatic field led to the concept of electric potential and potential difference. The properties of closed surface integral of electrostatic field led to Gauss' theorem which considerably simplified the calculation of electric fields in some special situations. The open surface integrals of magnetic fields were used in chapter 7 to describe electromagnetic induction.

APPENDIX B

TABLE OF PHYSICAL CONSTANTS

Constant	*Symbol*	*Value*
Charge of electron or proton	e	1.600×10^{-19} C
Rest mass of electron	m_e	9.110×10^{-31} kg
Rest mass of proton	m_p	1.6726×10^{-27} kg
Rest mass of neutron	m_n	1.6749×10^{-27} kg
Atomic mass unit	u	1.6604×10^{-27} kg
Universal gravitational constant	G	6.673×10^{-11} Nm^2 kg^{-2}
Mass of earth	M_e	6.000×10^{24} kg
Universal gas constant	R	8.314 J K^{-1} mol^{-1}
Avogadro constant	N	6.022×10^{23} mol^{-1}
Boltzmann constant	k	1.380×10^{-23} JK^{-1}
Mechanical equivalent of heat	J	4.186 J cal^{-1}
Stefan Boltzmann constant	σ	5.670×10^{-8} W m^{-2} K^{-4}
Speed of light in vacuum	c	2.998×10^{8} ms^{-1}
Planck's constant	h	6.626×10^{-34} Js
Permittivity of free space	ϵ_0	8.854×10^{-12} Fm^{-1}
Permeability of free space	μ_0	1.257×10^{-6} Hm^{-1}
Faraday	F	9.6487×10^{4} C mol^{-1}

ANSWERS

CHAPTER 1

C. Multiple-Choice Questions

1. (a), (c) **2.** (b), (d) **3.** (b) **4.** (a)
5. (a), (b), (c) **6.** (b) **7.** (a) **8.** (d)

D. Numerical Problems

1. 9.1×10^{-2} N **2.** 2.6N, 37° with x-axis **3.** 1.6g **4.** 2.4×10^{-8}s 1.3×10^{7} ms^{-1} **5.** $\pm6.5\times10^{-9}$ C **6.** 2 electrons **7.** 3.85×10^{-5} C, 1.15×10^{-5} C **8.** 3.6×10^{4} NC^{-1} towards the empty corner. **9.** -0.5μC **10.** $\pm1.0\ \mu$C, $\mp3.0\ \mu$C **11.** 5.6×10^{-11} C **12.** (a) The bigger charge produces a field of 1.8×10^{3} NC^{-1} at the site of the smaller charge, the smaller charge produces a field of 7.2×10^{2}NC^{-1} at the bigger charge (b) 1.44×10^{-4} N (repulsive) **14.** 1.0×10^{5} NC^{-1} pointing vertically up **15.** 2×10^{-7} NC^{-1}, vertically upward **16.** 2.4×10^{-16} N upward, **17.** 2.0 **18.** 1.375 m **19.** 0.0624 N away from triangle perpendicular to opposite side.

CHAPTER 2

C. Multiple-Choice Questions

1. (a) **2.** (c) **3.** (c) **4.** (d) **5.** (b) **6.** (d)
7. (b) **8.** (b) **9.** (d) **10.** (b) **11.** (b) **12.** (a)
13. (d) **14.** (b) **15.** (d) **16.** (c)

D. Numerical Problems

1. 1.7×10^{7} V **2.** 6.3×10^{4} V **3.** 1200 V **4.** 900 V **5.** 9000 V, 9×10^{4} Vm^{-1} **6.** 3600 V **7.** 3.7×10^{6} ms^{-1} **8.** 2880V **9.** -27.2 eV **10.** 6.0×10^{-15} m **11.** -0.9 J **12.** 1.1×10^{17} Vm^{-1} **13.** (a) 3.2×10^{-13} J, (b) 1.6×10^{-13} J, (c) proton. **14.** 1.0×10^{-5} C, 2.5×10^{-4} J **15.** 9×10^{-4}C, 1.8×10^{-3} C, 0.135 J, 0.27 J **16.** 6×10^{-4} C, 6×10^{-4} C, 0.06J 0.03J **17.** $Q_1 = 2.0\times10^{-4}$ C, $Q_2 = 7.7\times10^{-4}$ C, $V_1 = V_2 = 96$ V **18.** $Q_1 = Q_2 = 0$, $V_1 = V_2 = 0$ **19.** 43 pF **20.** (a) 160 V, (b) 0.14 J, (c) 0.128 J **21.** (a) 1.8×10^{-8}C, 4.4×10^{-6} J, 1.0×10^{5} Vm^{-1}, (b) 2.5 **22.** 500 μF **23.** $Q/4\pi\epsilon_0(R+r)/(R^2+r^2)$ **24.** (a) $Q_1 = Q_3 = 9\ \mu$C, $Q_2 = Q_4 = 16\ \mu$C (b) $Q_1 = 8.4\ \mu$C $Q_2 = 16.8\ \mu$C, $Q_3 = 10.8\ \mu$C, $Q_4 = 14.4\ \mu$C **25.** 4μF **26.** (a) $C_0 = 8.9$ pF, (b) $Q = 8.9\times10^{-10}$ C, (c) $E_0 = 1.0\times10^{4}$ Vm^{-1}, (d) $E = 1.4\times10^{3}$ Vm^{-1}, (e) $V = 57$ V, (f) $C = 16$ pF.

CHAPTER 3

C. Multiple-Choice Questions

1. (b) **2.** (b) **3.** (b) **4.** (d) **5.** (d)
6. (a) **7.** (b) **8.** (c) **9.** (b) **10.** (a)
11. (d) **12.** (b) **13.** (a) **14.** (c) **15.** (a)
16. (c) **17.** (c) **18.** (c) **19.** (c) **20.** (a)
21. (b) **22.** (b)

D. Numerical Problems

1. 1.6A, from B to A **2.** (i) 600°C, (ii) 3.75×10^{21} electrons **3.** 5.83×10^{28} **4.** (i) 5.2×10^{-4} ms^{-1}, (ii) 1.94×10^{3} s **5.** 3×10^{-14}s **6.** 3×10^{-6} cm **7.** 20Ω **8.** 1Ω **9.** (i) 0.25A, (ii) 0.5 V, 0.75 V **10.** 1.5 V, 0.5 Ω **11.** 1/273 K^{-1} **12.** 250°C, Yes **13.** 1/3Ω **14.** (i) 200Ω, (ii) 0.23V **15.** (i) 10Ω, (ii) 14Ω, (iii) 10Ω, (iv) 10Ω **16.** (a) two rows of 12 cells in series, (b) (i) 0.375A, (ii) 9 V **17.** 0.01Ω **18.** (a) From positive to negative terminal of the battery, (b) 2.5 V. **19.** 22.5 V **20.** (a) 3.5A, (b) 7V **21.** 13.5V **22.** 2000Ω, 1000Ω **23.** 0.25A, 7.5V **24.** 10/38 A, 35/38A, 45/38 A, 45/38 A **25.** (a) 1.53 V, (b) 60.0 cm **26.** (a) +33V, +32V, (b) A to B, (c) 22Ω

CHAPTER 4

C. Multiple-Choice Questions

1. (b) **2.** (b) **3.** (b) **4.** (d) **5.** (a)
6. (a) **7.** (d) **8.** (c) **9.** (c) **10.** (d)
11. (b) **12.** (d) **13.** (b) **14.** (c) **15.** (d)
16. (b) **17.** (d)

D. Numerical Problems

1. 10 Ω **2.** 200V **3.** 14 min **4.** 1 h 10 min 39 s **5.** (i) 1210 Ω (ii) Rs. 2.16 **6.** (i) 5 min 37.5 s (ii) 24 min **7.** 0.4 °C **8.** 400 °C **9.** 8 min 20s **10.** 3 h 20 min **11.** 1.0 Ω **12.** 5 h 47 min 8 s **13.** 75000 J **14.** 6.35 Ω **15.** 2.5 W **16.** 2.5 Ω, 7% **17.** (a) 1.5×10^{-2}V (b) 25V **18.** 820 °C **19.** 1.0 μV **20.** 1.53 V

CHAPTER 5

C. Multiple-Choice Questions

1. (c) **2.** (d) **3.** (a) **4.** (b) **5.** (c)
6. (c) **7.** (d) **8.** (c) **9.** (c) **10.** (d)
11. (d) **12.** (c)

D. Numerical Problems

1. (i) 1.26×10^{-2} T (ii) 1.12×10^{-3} T **2.** 2.45×10^{-2} T perpendicular to the wire **3.** 1.4×10^{-3} T directed into the page, normal to it **4.** 1.91 m **5.** 3.7×10^{-3} V **6.** 1.2×10^{-5} Nm^{-1}, repulsive **7.** 9.6×10^{-20} N **8.** 4×10^{-4} N along $+\ x$-direction **9.** (a) 3.0N, (b) 2.12 N, (c) zero. **10.** 6 cm from the 12A wire **11.** 3.75×10^{-24} N **12.** (a) 3.2×10^{-16} N parallel to current, (b) 3.2×10^{-16} N radially outward if the velocity is parallel to current, (c) zero **13.** 3.3 cm **14.** 2×10^{5} ms^{-1} **15.** 0.5 T **16.** 9 8 A **17.** 3×10^{-3} Nm **18.** 1.9*r* **19.** (a) 0.16 N, 0.06N, 8.3×10^{-3} Nm (b) 0.16 N, 0.104 N, 4.8×10^{-3} Nm (c) same torque **20.** (a) 1.14×10^{-3}T, into the page (b) 1.57×10^{-8} s

CHAPTER 6

C. Multiple-Choice Questions

1. (b)	**2.** (b) and (d)	**3.** (a)	**4.** (b)	**5.** (a)
6. (d)	**7.** (a)	**8.** (b)	**9.** (a)	**10.** (c)
11. (a)	**12.** (d)	**13.** (b)	**14.** (a)	**15.** (c)
16. (c)				

D. Numerical Problems

1. 1.0×10^{-2} T **2.** 0.032 Nm **3.** 0.064 Nm **4.** (i) 5.0 cm, (ii) 6.3 cm **5.** 5.76T **6.** (a) 1.05 G (b) 12.6 cm **7.** 15° with $\mathbf{B}_1$ **8.** 2000 **9.** 2.4 Nm **10.** 10 JT^{-1}

CHAPTER 7

C. Multiple-Choice Questions

1. (b)	**2.** (d)	**3.** (a)	**4.** (b)	**5.** (d)	**6.** (d)
7. (b)	**8.** (a)	**9.** (c)	**10.** (b)	**11.** (b)	**12.** (d)

D. Numerical Problems

1. (i) 2 V, (ii) 0.2 A **2.** (a) 50 V (b) zero; (c) 39.3 V **3.** 2.0 Wb **4.** (a) 0.87 V, (b) 1.37 V (c) (i) 1.2 m V, (ii) 7.2×10^{-8}N, (d) 369 mV: **5.** (a) (i) 251.3 V, (ii) $\mathcal{E} = 251.3 \sin 40\pi t$ (iii) zero, (b) (i) 300, (ii) 50 Hz, (iii) $\pi/4$, (c) $\tau = 4\sin^2 100\pi t$ **6.** (a) 0.5 mm (b) (i) 62.5 μC; (ii) 31.25 μC; **7.** (a) (i) 1.1 V (ii) 44 kV; (iii) 19.36 kW; (b) (i) 500 (ii) 1.01 A; (c) (i) 45.5 (ii) 0.11 **8.** (a) 1.28 mH, (b) Yes **9.** (a) (i) 8 J (ii) 0.05 H, (b) (i) 2.1 mH (ii) 51.8 μH, **10.** 0.53 V

11. (a) $\mathcal{E} = -M(\mathrm{d}I/\mathrm{d}t) = -M\omega\cos\omega t = -45.7\cos 100\pi t$ volt

(b) $\mathcal{E}_{\max} = 15.7$ V

CHAPTER 8

C. Multiple-Choice Question

1. (b) 2. (a) 3. (d) 4. (b) 5. (b)
6. (c) 7. (d) 8. (b) 9. (b) 10. (c)
11. (d) 12. (a) 13. (d)

D. Numerical Problems

1. (a) (i) 2A, (ii) 0.2 s, (iii) 10 As^{-1}, (iv) 2 J; (b) (i) 0.78 A, (ii) 6.06 As^{-1} (iii) 7.87 W, (iv) 3.1 W, (v) 4.77W; (c) (i) 0.27 A, (ii) 1.96 J. 2. (a) (i) 160 μF, (ii) 6.2 s, (iii) 38.8 kΩ (iv) 64%, (b) (i) 62.5 Ω, (ii) 86.5% 3. (a) 0.84 H (b) 41.4° 4. (a) 12 Ω, (b) 1.59 H, (c) 89°, (d) 1.2 Hz 5. (a) 11.7 V, (b) 219.7 V, (c) 3° (d) 2002.8 Ω (e) $I = 0.16 \sin (314t + \pi/60)$, (f) 0.11 A 6. (i) 257 Ω, (ii) 0.86 A, (iii) $V_R = 171$ V, $V_L = 273$ V, (iv) 39° 7. $V_1 = 183$ V, $V_2 = 57$ V, $V_3 = 147$ V 8. (a) 24 V, (b) 96 V, (c) zero, (d) 64 Hz

CHAPTER 9

C. Multiple-Choice Questions

1. (b) 2. (a), (b) 3. (d) 4. (d) 5. (a), (b) 6. (c)

D. Numerical Problems

1. 0.07 A 2. 2.8×10^{-7} T 3. $I_d = 1.5 \times 10^{-8}$ A opposite to the direction of electric field, **B** between plates is zero at all points 4. 5.0 μA 5. 0.15 A 6. 6.5×10^{-12} T 7. 2.25 8. (a) 1.5×10^{3} Vm^{-1} (b) 1.0×10^{-5} Jm^{-3}

CHAPTER 10

C. Multiple-Choice Questions

1. (c) 2. (b) 3. (c) 4. (d) 5. (a)
6. (d) 7. (d) 8. (b) 9. (d) 10. (d)

D. Numerical Problems

1. 5.2×10^{6} ms^{-1} 2. 4500 Å, 3750 Å 3. 1.5 4. 18 cm 5. 5.5×10^{-7}m 6. 5.5 mm 7. (a) 0.45 m, (b) 0.6 mm, (c) 0.75 mm 8. (a) 0, (b) $K/4$ 9. 2.4×10^{-3} mm 10. 8 11. 1.0 mm 12. 1.225×10^{-5} cm^2, 0.082 Wm^{-2}

CHAPTER 11

C. Multiple-Choice Questions

1. (c) **2.** (c) **3.** (b) **4.** (c) **5.** (d) **6.** (a)
7. (b) **8.** (a) **9.** (c) **10.** (c) **11.** (b) **12.** (d)
13. (a) **14.** (c)

D. Numerical Problems

1. 70° 2. 5.7° 3. 82.3 cm, 5m 4. 1.2 cm 5. 26.4 cm
6. 37.5 cm 7. 1 ms^{-1} **8.** 114° 9. 60 cm **10.** 1.41 **11.** 1.53
14. 60 cm **15.** 19.2 cm **16.** (a) 37.5 cm, 7.5 cm (b) 50 cm, 10 cm
17. 1.4 cm **18.** 30 cm, −60 cm **19.** 2.7 **20.** 5, 120 cm, 1 cm.

CHAPTER 12

C. Multiple-Choice Questions

1. (a) 2. (b) **3.** (a) 4. (b) 5. (d) 6. (b)
7. (a) **8.** (c) **9.** (d) **10.** (c)

D. Numerical Problems

1. 0.02 m 2. 1.8×10^{-11} C kg^{-1} 3. 1.12×10^{-4} m, 7.28×10^{-18} C
4. 2.9 eV 6. 0.69 V, no photo-emission from the second tube 7. 2.5 V
8. 6000 : 1 9. 1.99×10^{-18} J **10.** 1.99×10^{-6} m **11.** (a) 1.21×10^{15}Hz
(b) 2.48×10^{-7} m (c) 8.56×10^{-19} J, (d) 5.35 V **12.** 1.89×10^{-18} J
13. 5.93×10^{5} ms^{-1}, 1.71×10^{-19} J

CHAPTER 13

C. Multiple-Choice Questions

1. (d) 2. (b) 3. (a) 4. (d) 5. (d) 6. (b)
7. (c) 8. (a) **9.** (b) **10.** (a) **11.** (c) **12.** (a)
13. (d) **14.** (c) **15.** (a)

D. Numerical Problems

1. 106 Å 2. 0.85 eV, 3.4 eV, 6.16×10^{14} Hz, 4.87×10^{-7} m, yes
3. 6.57×10^{-7} m, yes **4.** 0.058 m 5. 25.2 g 6. 4.23 MeV 7. 100 days
8. 1.88×10^{4} s **9.** 1.4×10^{4} years **10.** (a) 1.99 years, (b) 9.97×10^{5} Years
11. 6.56×10^{-15} cm^{3} **12.** 0.6 Ci **13.** 2.60×10^{-21} per gram, 0.11 Ci,
15. 0.36 kg, **16.** 8.78 days **17.** 26.7 MeV, 1.56×10^{-12} kg s^{-1} **18.** 3.76×10^{8} N

CHAPTER 14

C. Multiple-Choice Questions

1. (c) 2. (c) 3. (d) 4. (d) 5. (b) 6. (a)
7. (d) 8. (a) 9. (c) 10. (c) 11. (b) 12. (a)
13. (a) 14. (b) 15. (d)

D. Numerical Problems

1. (a) 980, (b) 0.04 mA, 1.96 mA 2. 1000 3. 0.8 mA, 39.2 mA
4. (a) 727, (b) 706, 5. 24 mA

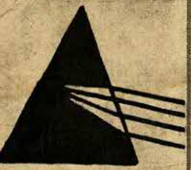
CHAPTER 15

C. Multiple-Choice Questions

1. (b) 2. (d) 3. (d) 4. (c) 5. (a)

D. Numerical Problems

1. 5.977×10^{24} kg 2. 7.45×10^{22} kg, 5.97×10^{24} kg 3. 1.998×10^{30} kg
4. 5.976×10^{2} W m^{-2}, 320.4 K 5. 1.388×10^{3} Wm^{-2} 6. 3.844×10^{8} m
7. 5796 K 8. 1: 10^{4} 9. 35.7″ of arc. 10. 2.4×10^{41} J



FOR CLASS XII

PHYSICS

The book is written strictly according to the latest CBSE syllabus to be implemented in 1989-90.

Salient Features

- Each concept in a topic is explained in a comprehensive manner with a number of illustrations.
- Stress has been laid on practical application throughout the book.
- Each concept is followed by a number of solved examples and solved numericals.
- Exercises include descriptive and short-answer questions followed by a number of numerical problems.
- Answers to numerical problems are provided at the end of the book.

Dr. N K Bajaj has been teaching physics at St. Stephen's College, University of Delhi, for the last three decades. Between 1969 and 1971, he was a member of the physics faculty at the University of Maryland, USA. He has been a resource person in physics at various workshops organised by the NCERT for the improvement of teaching and evaluation in physics in schools.

Other Books of Interest
Kapoor : *Chemistry for Class XII*
Sankaran : *Biology for Class XII*

Tata McGraw-Hill Publishing Company Limited
4/12 Asaf Ali Road, New Delhi 110 002

0-07-460044-3